BIT

生物信息学

汇报人：陈籽旭 刘天依 余宏骏 乔江洋 张洪洋

时间：2024-11-6

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

德以明理 学以精工

BIT

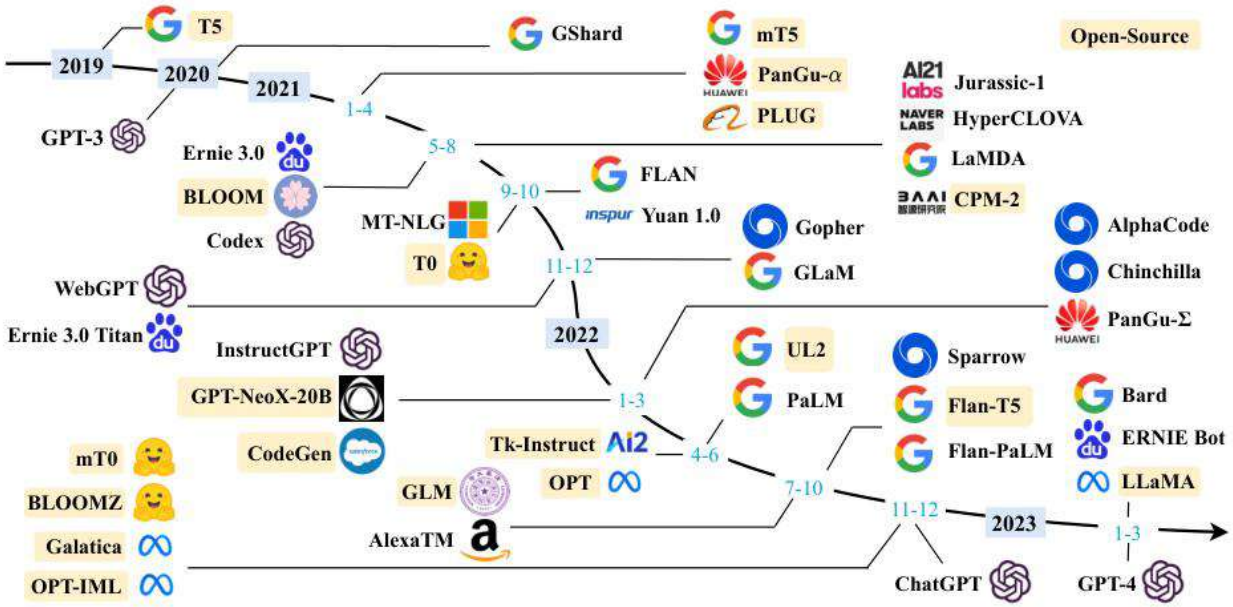
计算机科学与技术前沿——生信方向专题汇报

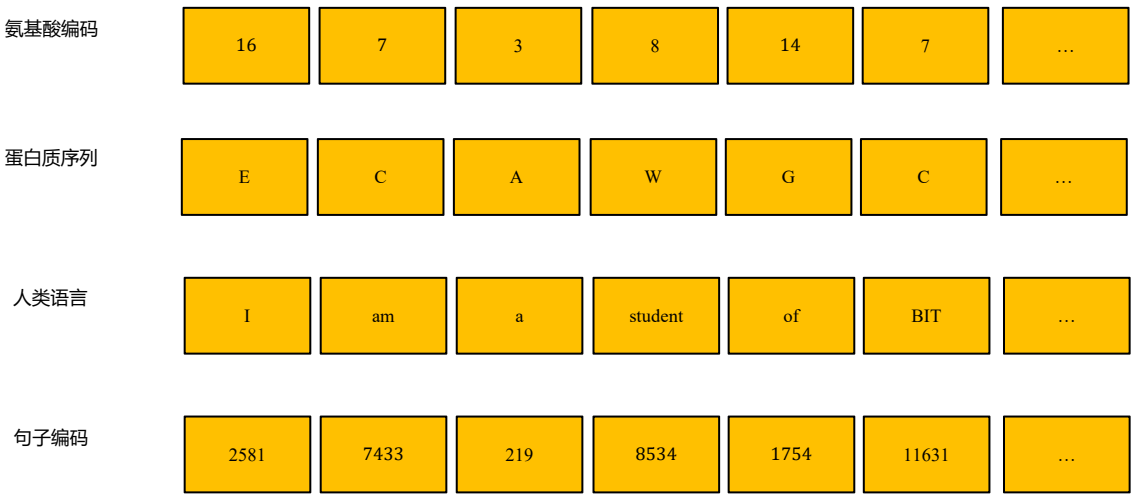
课题背景及方法论

汇报人：陈籽旭　　时间：2024-11-6

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

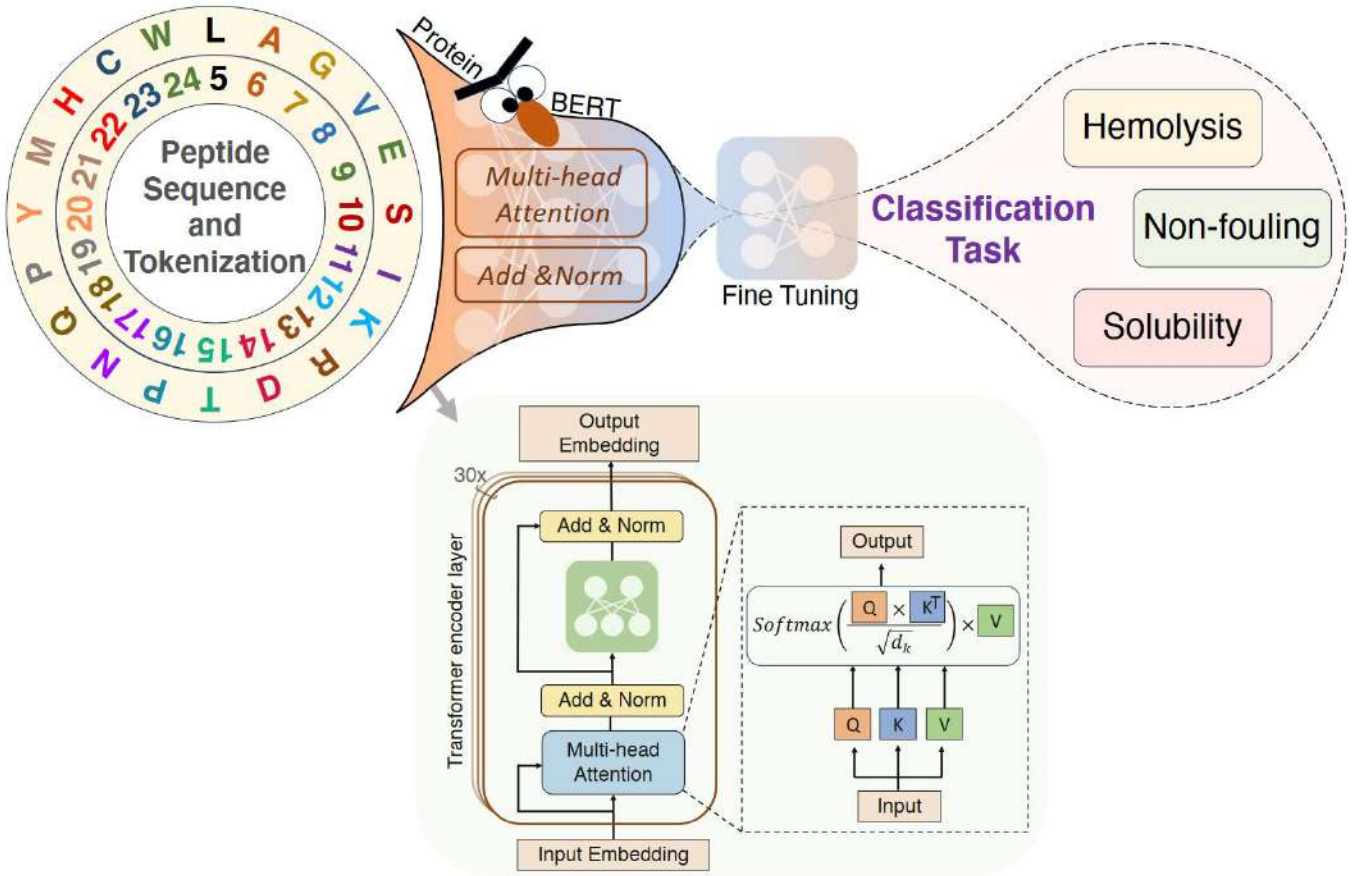
德以明理　学以精工





- 大量蛋白质被发现，人工注释蛋白质成本高，速度慢
- 自然语言处理发展，大模型在各类任务中表现良好
- 蛋白质序列可以理解为自然语言中的句子，氨基酸理解为单词
- 可以通过语言模型捕捉氨基酸之间的关系

PeptideBERT: A Language Model based on Transformers for Peptide Property Prediction



1. 语言模型发展->蛋白质序列预测（表示为文本）
2. 目标：预测肽的三个特性，溶血，溶解度，非污垢(hemolysis, solubility, non-fouling)
3. 氨基酸有机分子（organic molecules containing amino acids)，长度和排列决定了蛋白质的结构和生物特性。

```
m1 = [
    '[PAD]','A','R','N','D','C','Q','E','G','H',
    'I','L','K','M','F','P','S','T','W','Y','V'
]
m2 = dict(zip(
    ['[PAD]','[UNK]','[CLS]','[SEP]','[MASK]','L',
    'A','G','V','E','S','I','K','R','D','T','P','N',
    'Q','F','Y','M','H','C','W','X','U','B','Z','O'],
    range(30)
))
```

20种氨基酸，使用数组索引标记（从1开始）[A, R, N, D, C, Q, E, G, H, I , L, K, M, F, P, S, T, W, Y, V]; ([PAD]: 补全字符, [UNK]: 低频词或未在词表中出现的词, [CLS]: 开头，后面跟序列，分类标签, [SEP]: 两个句子之间的分隔符, [MASK]: 填充被掩盖的字符)

对于蛋白质数据集的数据增强，提高分类精度：

1. 随机替换
2. 随机删除
3. 用A随机替换
4. 随机交换
5. 随机插入A
6. 随机掩码：[MASK]

Table 1: Ablation results for different augmentation techniques for *Solubility* prediction. Baseline accuracy (without any augmentations) is 69.175%

Augmentations applied	Train set size	Accuracy(%)
random_replace(2%)	29892	68.694
random_delete(2%)	29892	68.814
random_replace_with_A(2%)	29892	68.573
random_swap(2%)	29892	70.018
random_insertion_with_A(2%)	29892	69.597
random_swap(2%), random_insertion_with_A(2%)	44838	68.453
random_swap(2%), random_insertion_with_A(1%)	44838	68.814
random_swap(3%)	29892	68.814
random_replace_with_A(2%), random_insertion_with_A(2%)	44838	69.054

随机交换表现最好，准确率提高0.843%

为了可视化PeptideBERT对肽序列的理解和分类能力，从隐藏状态提取CLS标签，并使用t分布随机近邻嵌入(t-distributed Stochastic Neighbor Embedding)t-SNE进行可视化，评估高维空间的相似性，吸引相似数据点，排斥不相似点。大小为480的CLS标签嵌入可视化，同一类同颜色。结果表明，模型仅仅根据序列信息对肽进行分类，CLS捕获了单个肽的特征。

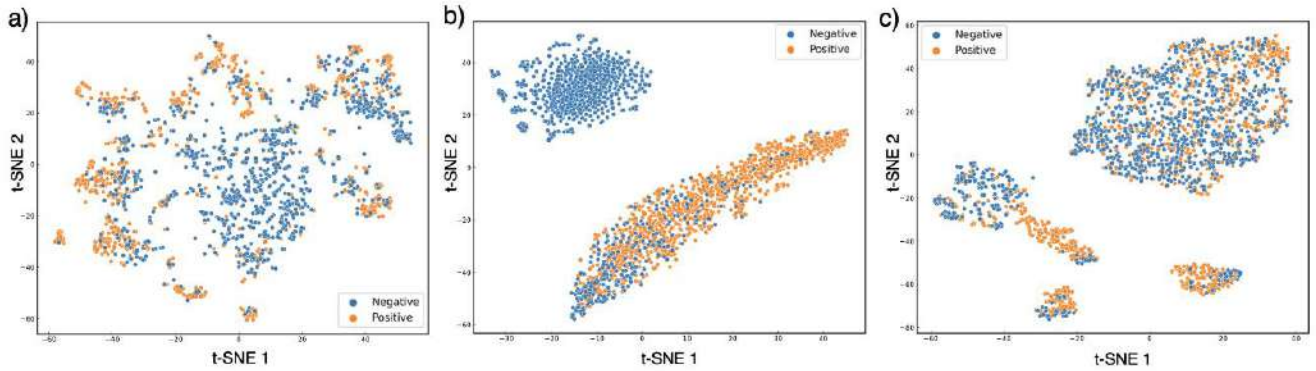
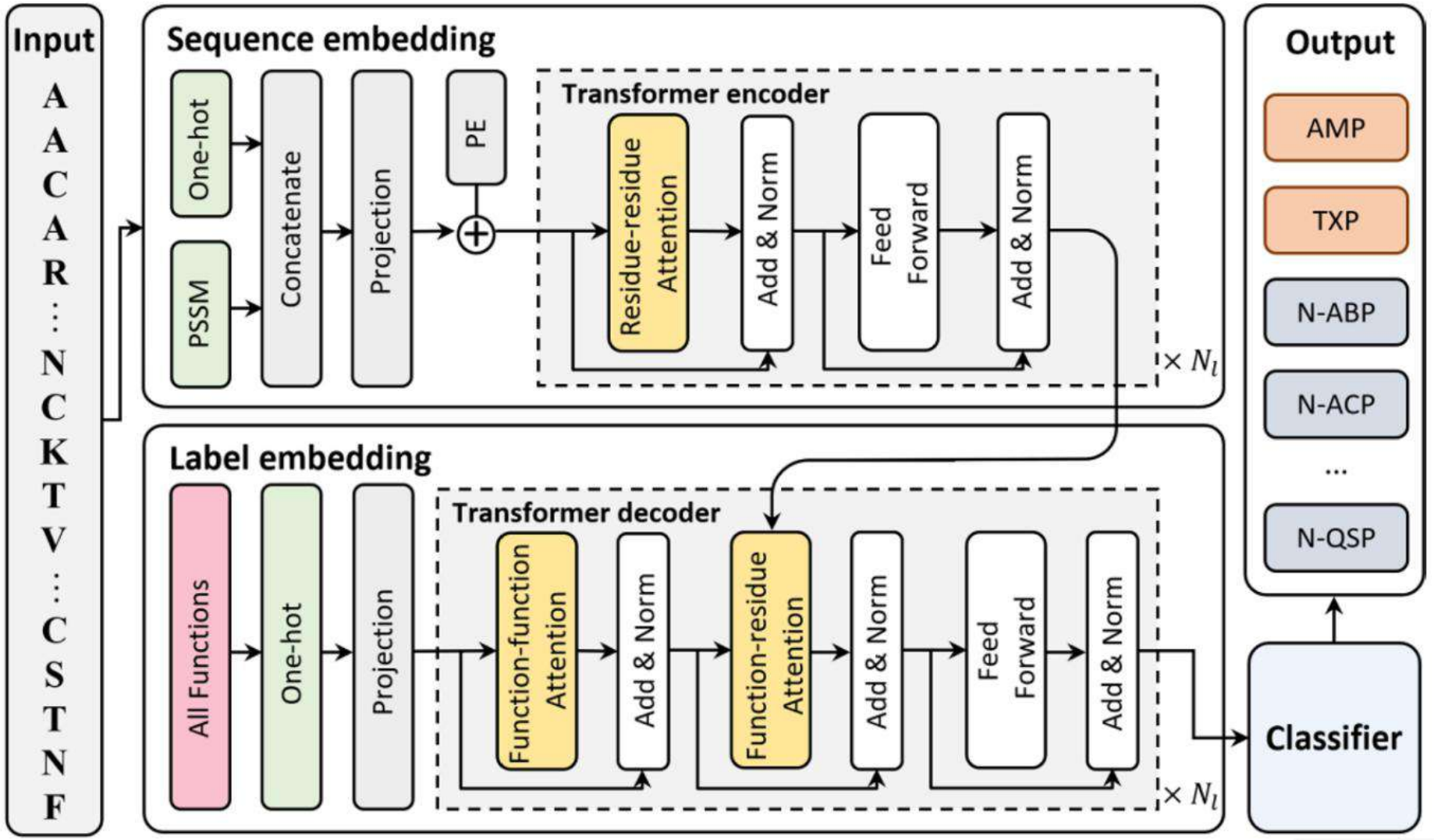


Figure 3: t-SNE visualization of peptide properties (a) Hemolysis, (b) Non-fouling and (c) Solubility. The [CLS] token embedding from the last hidden state of PeptideBERT is visualized after dimensionality reduction.


Input
A A C A R ... N C K T V ... C S T N F
Sequence embedding
One-hot
PSSM
Concatenate
Projection
PE
Transformer encoder
Residue-residue Attention
Add & Norm
Feed Forward
Add & Norm
× N_l
Label embedding
All Functions
One-hot
Projection
Transformer decoder
Function-function Attention
Add & Norm
Function-residue Attention
Add & Norm
Feed Forward
Add & Norm
× N_l
Output
AMP
TXP
N-ABP
N-ACP
...
N-QSP
Classifier

Last position-specific scoring matrix computed, weighted observed percentages rounded down,

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-1	-2	-2	-3	-2	-1	-2	-3	-2	1	2	-2	7	0	-3	-2	-1	-2	-1	1
2	S	2	-1	0	-1	-1	0	0	0	-1	-2	-3	0	-2	-3	-1	4	1	-3	-2	-2
3	L	-1	-2	-3	-4	-2	-2	-3	-4	-3	1	4	-2	5	0	-3	-2	0	-2	-1	0
4	V	-1	-3	-3	-3	-1	-2	-3	-3	-3	2	1	-2	3	-1	-3	0	0	-3	-2	4
5	I	-1	-2	-2	-2	-2	2	0	-3	-2	2	0	-1	5	0	-3	0	1	-2	-1	1
6	P	-1	-1	2	2	-3	2	2	-1	-1	-4	-4	-1	-3	-4	4	1	0	-4	-3	-3
7	E	0	-1	-1	3	-4	4	4	0	-1	-4	-4	0	-2	-4	-2	0	0	-4	-3	-3
8	K	-2	-1	2	3	-5	2	5	-3	-1	-4	-4	1	-3	-4	-2	-1	-2	-4	-3	-4
9	F	-3	-4	-4	-5	-3	-4	-4	-5	-2	2	1	-4	0	6	-4	-3	-3	-1	3	1
10	Q	-2	4	0	-2	-4	5	0	-3	-1	-4	-4	4	-2	-4	-3	-1	-2	-4	-3	-4
11	H	-1	-2	-2	-1	-4	-2	-2	-4	8	-4	-4	-3	-3	0	-2	-3	-3	-1	6	-4
12	I	-4	-3	-6	-7	-5	-4	-6	-5	-5	3	0	-5	10	1	-6	-5	-4	-5	-4	-2
13	L	5	-4	-4	-5	-3	-4	-4	-3	-5	0	0	-4	0	-4	-1	-1	-1	-5	-4	4
14	R	-5	8	-4	-5	-7	-3	-4	-6	2	-4	-6	-2	-5	-5	-6	-4	-5	-6	2	-6
15	V	-4	-6	-7	-7	-4	-6	-6	-7	-7	7	1	-6	-2	-2	-6	-6	-4	-6	-4	2
16	L	5	-3	-4	-4	0	-1	-3	0	-2	-2	1	-3	1	-2	-4	1	-2	-4	-2	-1
17	N	-3	-1	4	-1	-6	-2	-4	6	-4	-7	-7	-3	-6	-6	-5	-3	-3	-6	-6	-6
18	T	-1	-3	-3	-5	-4	-2	-4	-5	-5	3	-2	0	-2	-4	-5	-3	3	-5	-4	5
19	N	-3	-2	5	6	-6	-3	1	-2	0	-6	-5	-3	-5	-6	-5	-2	-4	-7	-5	-6
20	I	-4	-5	-6	-6	-4	-5	-6	-7	-6	5	4	-5	-1	-1	-6	-5	-4	-5	-4	2
21	D	-3	-5	0	4	-6	-3	-2	-5	-5	-6	-6	-2	-6	-7	8	-2	-4	-7	-6	-3
22	G	-1	3	2	2	-4	-1	-1	3	-3	-4	-3	1	-3	-5	-2	-1	1	-5	-4	-2
23	R	-2	0	6	2	-5	0	1	-2	4	-6	-6	2	-4	-5	-4	-2	0	-6	-4	-5
24	R	-4	1	-2	-4	-6	0	-1	-5	-4	-5	-4	7	-2	-6	-5	-3	-4	-6	-5	-5
25	K	-2	6	0	-4	-2	2	-2	-5	5	-4	-4	2	-4	-5	1	-2	-1	-5	-3	-4
26	I	0	-5	-5	-5	-2	-4	-5	-2	-5	4	1	-5	1	-3	-5	-3	0	-5	-4	5
27	A	-1	0	-2	0	-3	0	4	-4	-1	-1	-1	-1	1	0	1	-2	-2	0	0	3
28	F	-3	-2	-4	-5	-4	-2	-5	-6	-4	6	-1	-5	-1	2	-5	-2	-2	-1	4	2
29	A	6	-5	-4	-5	-4	-4	-2	2	-5	-5	-5	-4	-4	-6	-4	2	-3	-6	-5	-4
30	I	-5	-6	-7	-7	0	-5	-6	-7	-6	-1	6	-6	1	-3	-6	-3	-4	-5	-5	-2
31	T	0	0	-2	-4	0	2	-3	-5	-5	-4	-4	-1	-2	-6	-2	-1	7	-6	-5	-4
32	A	1	-3	-3	1	-3	-1	-1	-1	-1	-5	-4	0	-2	-1	-5	1	-3	-2	7	-4
33	I	-5	-7	-7	-7	-5	-6	-7	-7	-7	8	0	-6	2	-1	-6	-6	-4	-6	-5	2
34	K	-5	-1	-2	-2	-3	-4	-4	-5	5	-5	-4	2	-4	4	-4	-5	-4	-2	8	-3
35	G	-4	-7	-5	-5	-7	-6	-6	8	-6	-8	-8	-6	-7	-7	-6	-4	-6	-7	-7	-7
36	V	-5	-7	-7	-7	-1	-7	-7	-8	-7	8	-1	-6	-1	-4	-7	-6	-4	-6	-5	3
37	G	-4	-6	0	-5	-1	-6	-6	7	-6	-8	-8	-6	-7	-7	-6	-2	-5	-7	-7	-7
38	R	-3	5	0	-2	-3	-1	-2	-3	1	0	1	2	-3	-4	2	-2	-2	1	0	-2
39	R	0	2	0	-2	-4	-2	-3	-3	1	-4	-3	2	-2	-3	-1	3	4	-1	-2	-2
40	Y	-2	4	-1	-4	-4	-2	-3	-4	-1	-2	0	2	0	1	-4	1	3	-4	1	-2
41	A	6	-5	-4	-5	0	-4	-4	-2	-5	-1	-5	-4	-5	-6	-5	3	-3	-6	-5	-4

红框中的内容是一个L*20维的矩阵，L代表着蛋白质序列的长度（图中绿色部分），20代表着20中氨基酸（图中蓝色部分）

对于其中的数字，十分常见的一个解释是：对于PSSM中的一个元素P_{ij}，其数值表示序列第i个位置上的氨基酸在进化过程中突变成第j个氨基酸的可能性，若值为正，就表示可能性越大；反之则表示可能性越小。

BIT

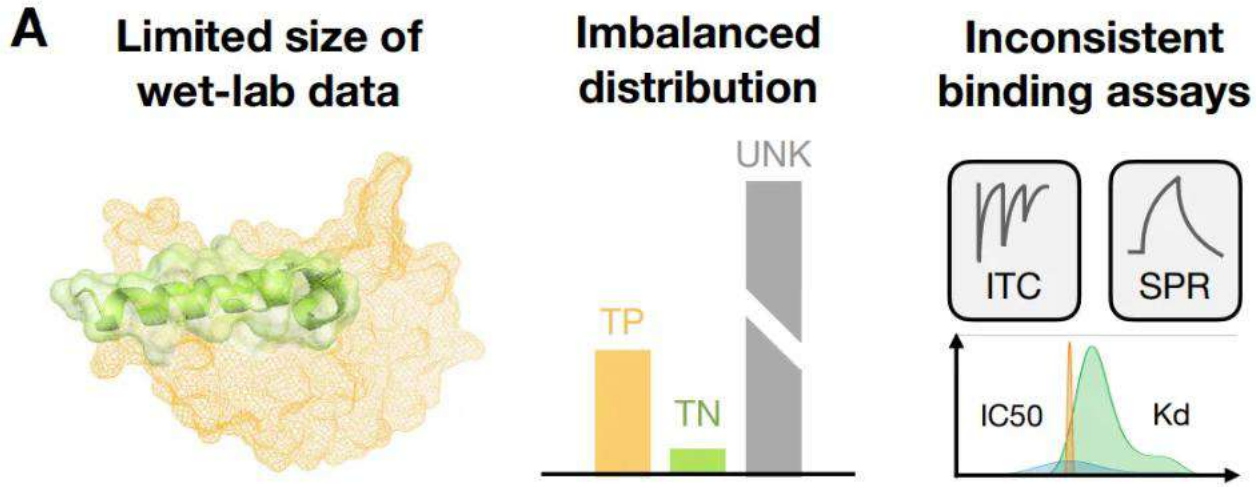
计算机科学与技术前沿——生信方向专题汇报

PDB-BRE：基于蛋白质数据库的配体-蛋白质相互作用结合残基提取器

汇报人：刘天依　　时间：2024-11-6

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

德以明理　学以精工



蛋白质数据库的重要性

资源通常只支持单一类型的配体，并且无法在分析新型复合物时获得满意的结果。

配体-蛋白质相互作用的研究价值

大多数研究都集中在分析配体-蛋白质相互作用，而忽略了插入和修饰残基的附加情境。

PDB-BRE的研发必要性

全面提取复合物中配体-蛋白质相互作用的结合残基。并且采用多进程处理机制，提高处理速度和效率。

PDB-BRE的数据来源与处理

数据来源

PDB-BRE使用PDB数据库作为主要数据来源，提取配体-蛋白质相互作用及其对应的结合残基。用户可以自动下载所需的PDB文件，也可以分析用户定义的PDB文件。

数据处理

PDB-BRE自动从PDB文件中提取必要的序列和残基坐标信息，进行筛选和映射。

数据整合

PDB-BRE支持将分析结果与UniProt数据库进行整合，以便研究人员可以更好地整合不同来源的数据。首先从PDBe数据库中提取分析的蛋白质链的标识，然后使用Needleman-Wunsch算法进行序列比对，最后根据最佳匹配映射蛋白质获得相应的残基标签。

PDB-BRE的基本原理和实现方法

01 数据收集与处理

PDB-BRE软件可以从PDB数据库中自动下载所需的PDB文件，也可以分析用户定义的PDB文件。从中提取配体与蛋白质相互作用的结合残基信息。

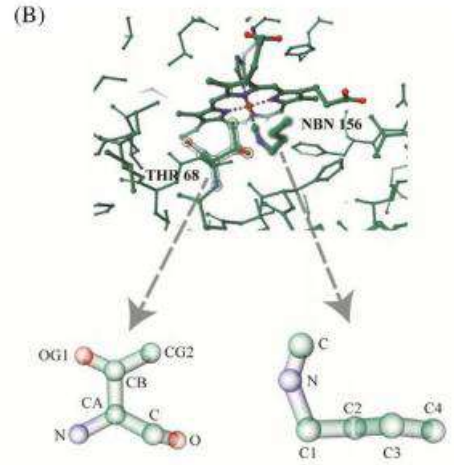
02 相互作用提取

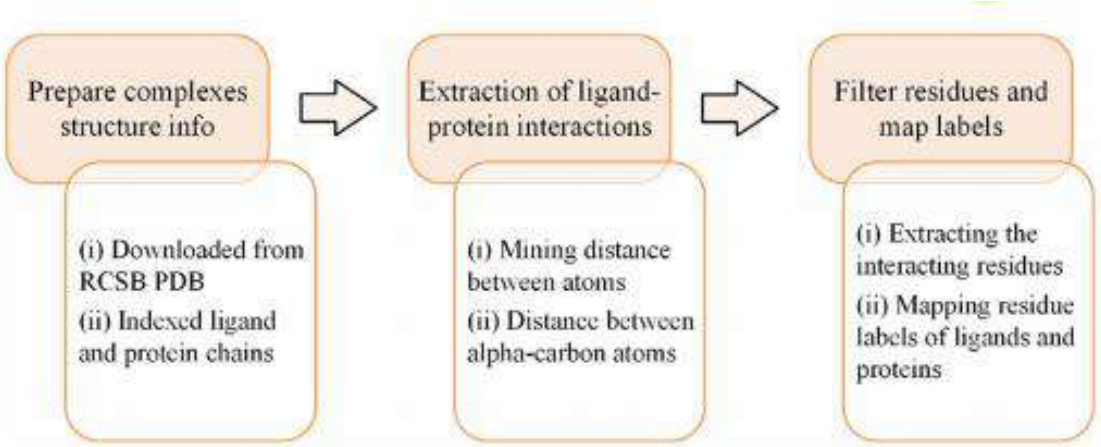
PDB-BRE确定每个配体和蛋白质链之间的相互作用。包括最小原子距离或α-碳原子距离。根据预设的距离阈值计算配体和蛋白质残基之间的结合位点标签。

03 过滤残基和映射标签

在处理插入残基时，PDB-BRE会保留特定的残基编号，以便更好地表示这些关联并促进结构比较。在处理修饰残基时，PDB-BRE支持将非标准氨基酸残基还原为标准氨基酸残基，从而为下游分析提供更准确的数据集。

04 多进程处理机制





PDB-BRE的结果分析

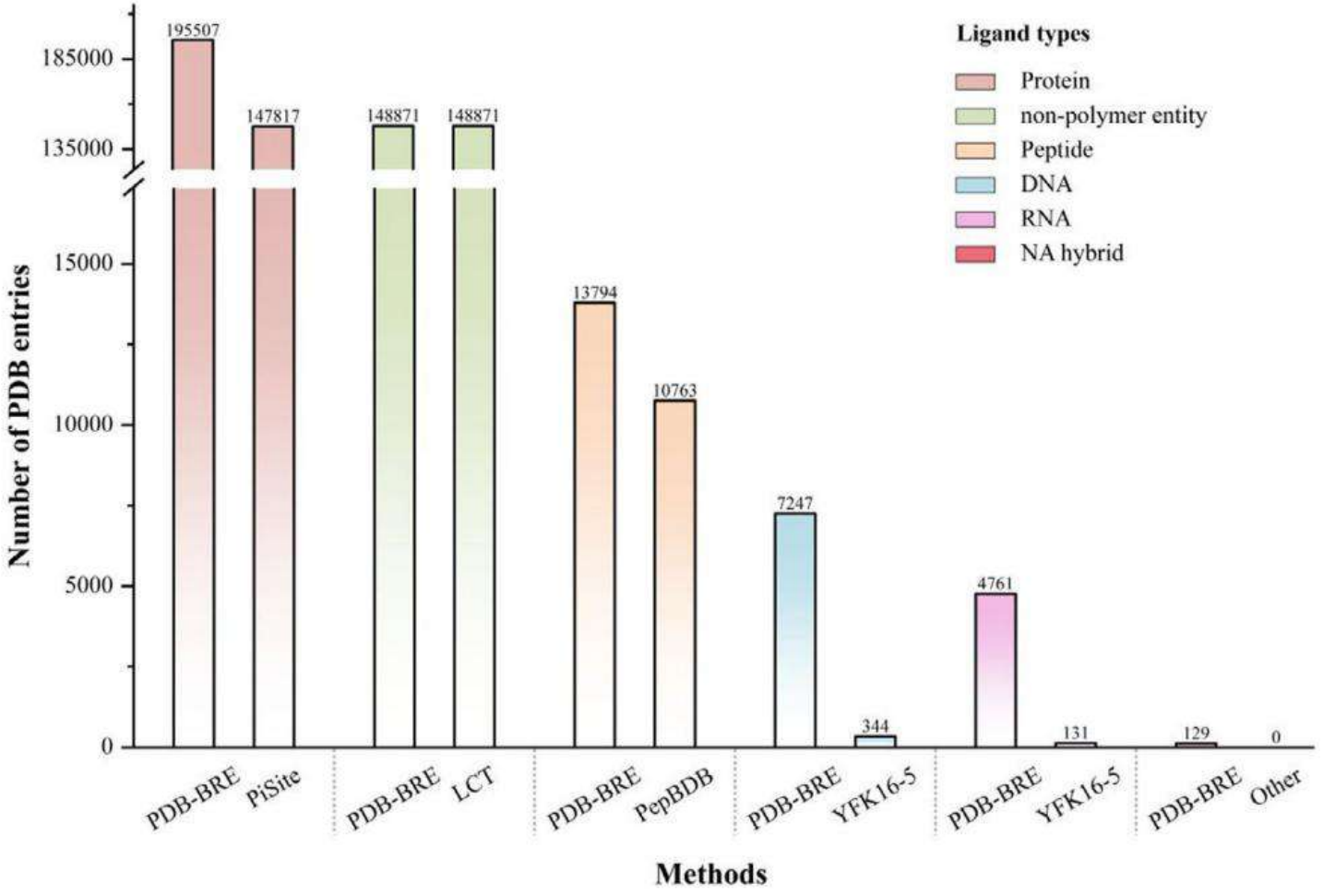


FIGURE 4 Comparison of the number of PDB entries supported by PDB-BRE and several other resources.

PDB-BRE的优势分析

准确性高

PDB-BRE能够处理插入和修改的残基，从而获得更全面的分析结果。

适用范围广

PDB-BRE还可以处理多种类型的配体，包括蛋白质、肽、DNA、RNA、DNA-RNA杂合物和非聚合物。

操作简便

PDB-BRE支持用户自定义和批量分析，使其比其它资源更灵活。。

效率高

BIT

计算机科学与技术前沿——生信方向专题汇报

用于多肽-HLA复合物建模的机器学习方法

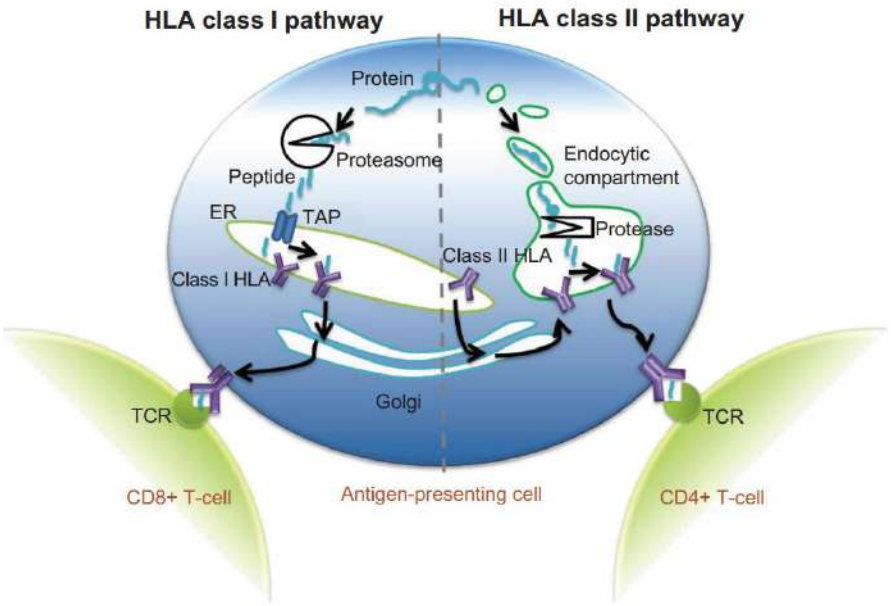
汇报人：余宏骏　　时间：2024-11-6

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

德以明理　学以精工

HLA（人类白细胞抗原）是一组与免疫系统相关的基因，它们编码一类被称为抗原呈递分子（MHC）的蛋白质。这些蛋白质在细胞表面展示“抗原”（即外来物质的片段或体内异常蛋白质片段）给免疫系统的T细胞，T细胞通过T细胞受体（TCR）识别HLA-多肽复合物，从而启动免疫反应。

当HLA向TCR呈递与自身肽结构相似的多肽时，可能会引发自身免疫反应。外源性药物可能会与抗原蛋白反应，插入到HLA的结合槽中，或干扰HLA-多肽-TCR复合物，进而导致不良事件。HLA、多肽和TCR的多样性都影响免疫反应，使得理解这些不良事件的潜在机制变得具有挑战性。然而，最新研究表明，通过考虑HLA结合槽内的结合肽，可以提高分子建模和预测的性能。因此，**理解HLA-多肽结合有助于解释药物与HLA之间的相互作用机制**。



- 针对预测HLA-多肽结合亲和力任务，目前三种主要的建模策略：

（1）多层感知机（MLP），如NetMHCIIpan-4.3；

（2）基于序列信息的模型，如基于Transformer的MHCAttnNet；

（3）**基于序列信息和结构信息的模型**，如基于卷积神经网络（CNN）的DeepMHCII，基于图神经网络（GNN）的Graph-pMHC。

- 从特异性角度可以将任务分为等位基因特异性和泛特异性两类：

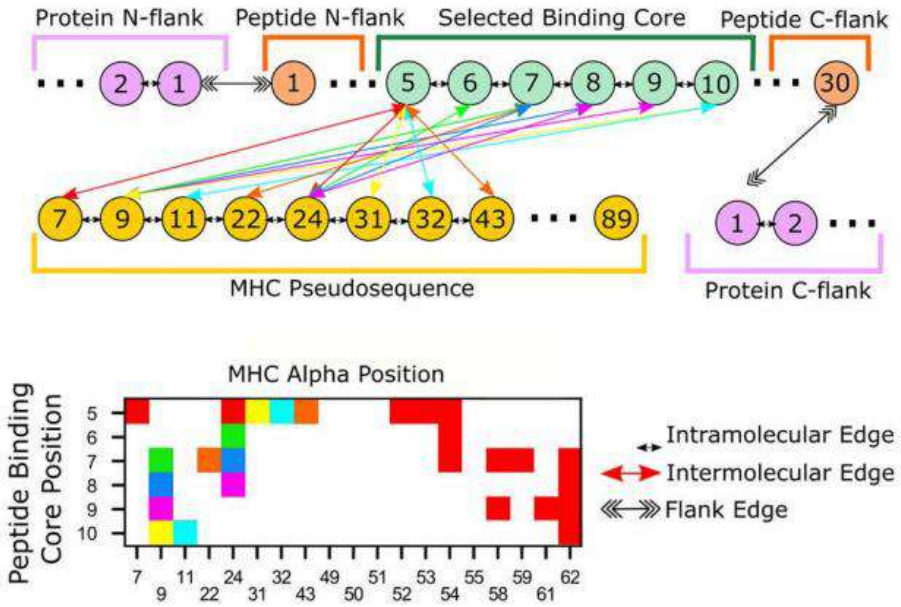
（1）等位基因特异性方法只能预测训练集中含有的等位基因的MHC-II分子的结合亲和力；

（2）**泛特异性方法**即使没有某些等位基因的MHC-II分子的训练数据也可以预测该等位基因MHC-II分子的结合亲和力。

现有研究与策略相关工作：Graph-pMHC

MLP、CNN等方法仅将多肽和MHCII视为一条氨基酸序列，而忽略了结构相互作用的重要信息。**图神经网络（GNNs）**通过利用连接并显示节点（残基）之间相互作用的边的先验知识，更好地捕捉真实的蛋白质系统。

本工作的图结构中，**通过锚定残基定义多肽的结合核心**，形成一个9-mer的结合核心。多肽结合核心和其他不与MHC接触的残基都包含在图中。**使用伪序列来表示MHC序列**，仅包括与多肽相邻的MHC残基。**利用图的边特征来表示残基之间的相互作用类型**，即分子间相互作用（多肽与MHC之间的相互作用）、分子内相互作用（多肽、MHC或侧翼序列内部的相互作用）以及蛋白质侧翼相互作用（肽与侧翼序列之间的相互作用）。通过将这些序列连接到各自独特的边标记，能够创建整个呈递通路的端到端模型。



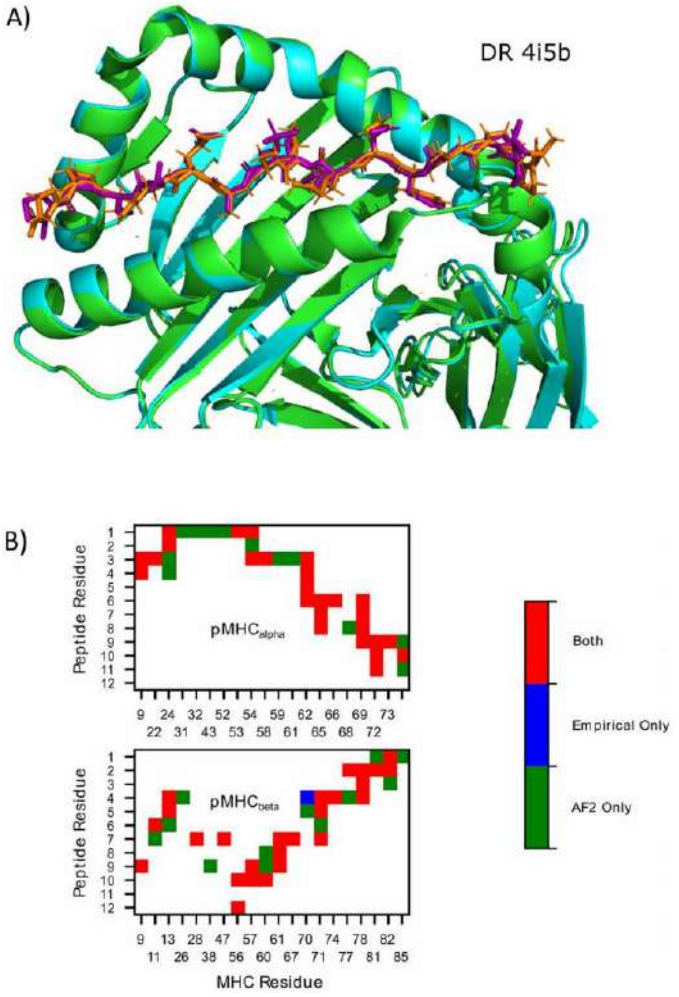
现有研究与策略相关工作：Graph-pMHC

（1）通过Alphafold2预测pMHCII邻接矩阵

HLA-DR、HLA-DP和HLA-DQ是不同的编码MHCII的基因，尽管在呈递抗原给T细胞的作用上并无差异，但它们在基因位点和多态性上有所不同，导致抗原呈递的变化，因此它们的pMHCII结构可能也会有所不同。

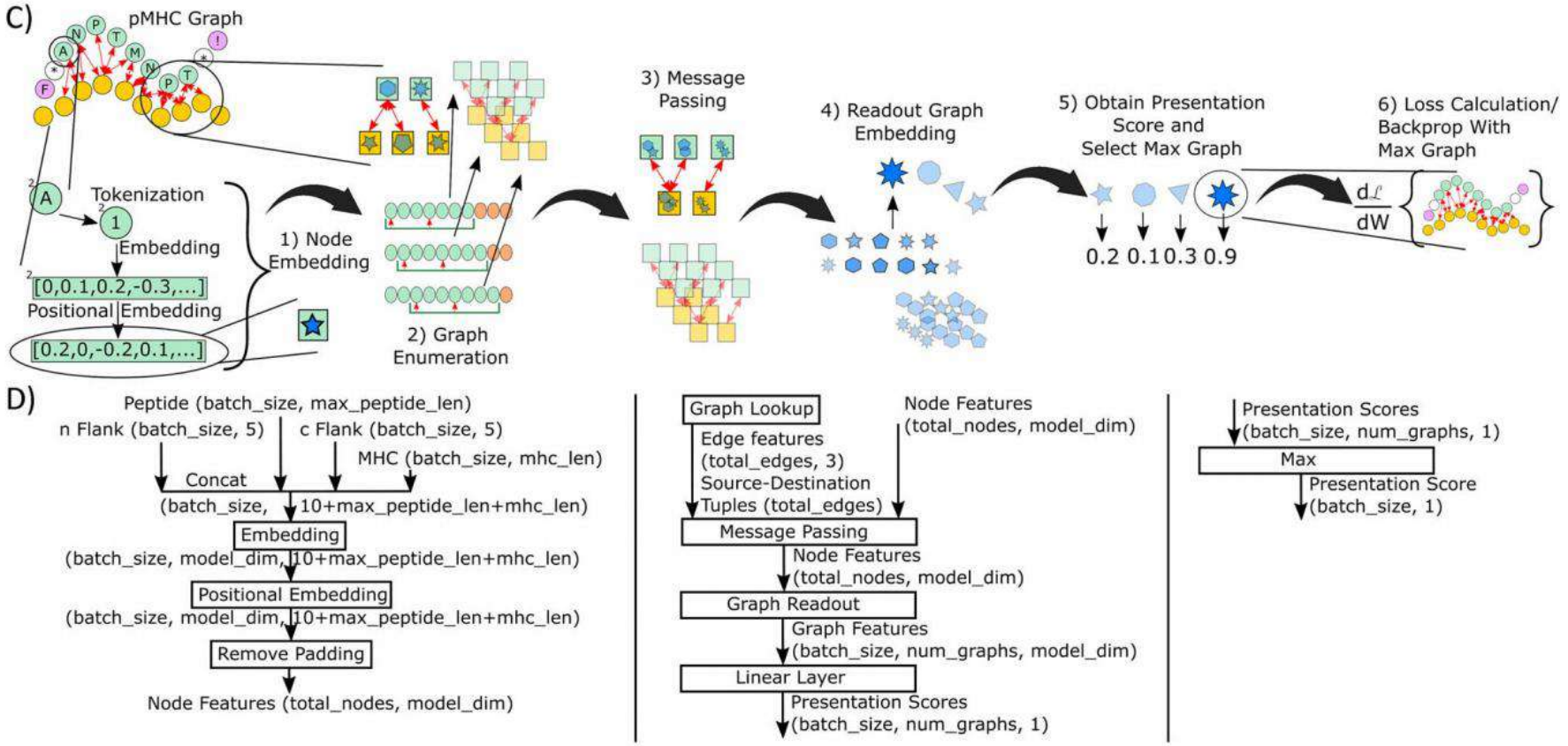
因此使用alphafold2-multimer来识别pMHCII的邻接矩阵，以预测pMHCII残基相互作用的图结构。

最终为每个基因（HLA-DR、HLA-DP和HLA-DQ）获得一个标准邻接矩阵。并且通过消融实验确定邻接关系在等位基因、多肽序列和多肽长度之间是保守的。



现有研究与策略相关工作：Graph-pMHC

（2）Graph-pMHC模型架构

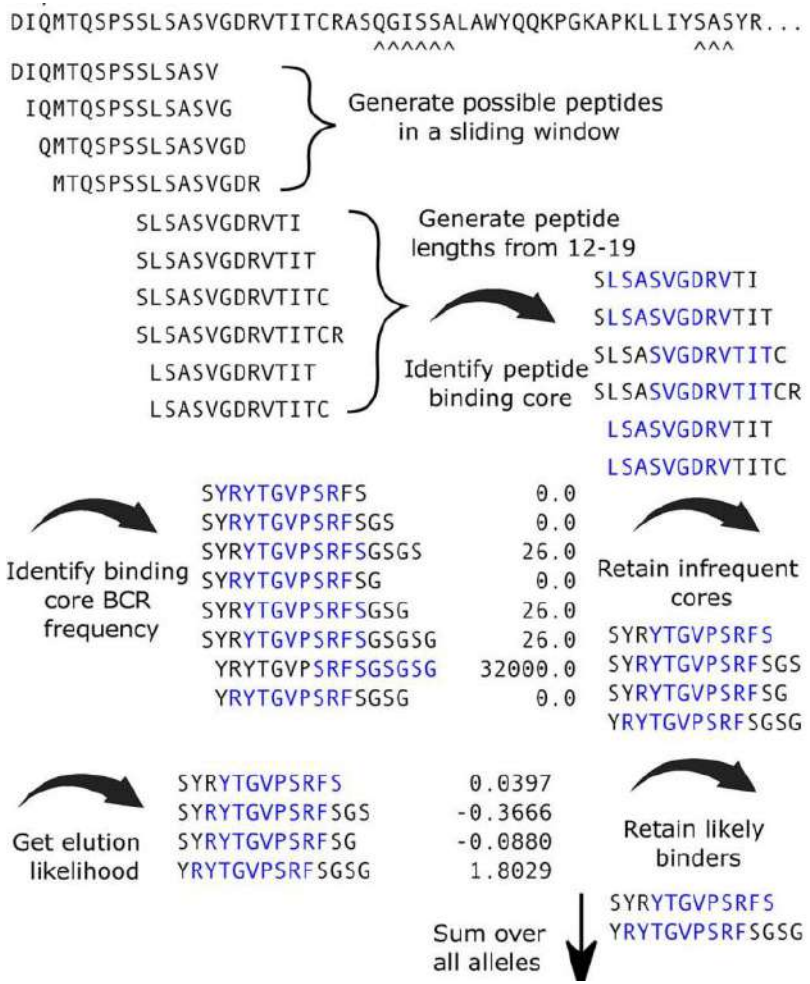


现有研究与策略相关工作：Graph-pMHC

（3）在抗体免疫原性上的应用

pMHCII模型的一个关键应用是对抗体药物进行去免疫原性。因此本工作开发了一种方法，可以从多肽的成分中评估抗体的免疫原性风险。

首先，**获取所有长度为12至19的多肽的呈递分数和结合核心**。接着，去除任何通过OASis数据库发现其结合核心存在于22个以上受试者中的多肽，并去除呈递logit分数低于0的多肽，并计算总肽的数量。这一过程对于选择的八种常见DR等位基因重复进行。最终，**总共的独特结合核心数量用于表示该抗体的免疫原性风险**。



BIT

计算机科学与技术前沿——生信方向专题汇报

利用官能团信息和超图结构预测蛋白质-蛋白质相互作用调节剂的层次图神经网络框架

汇报人：乔江洋

时间：2024-11-6

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

德以明理 学以精工

蛋白质-蛋白质相互作用（Protein-protein interaction，PPI）在信息传导、基因表达、细胞分裂等细胞活动中发挥着关键作用，将其作为治疗靶点进行调节是一种发现新型药物的策略。过去20年间，人们进行了大量实验研究，不断探索着**PPI调节剂（Protein-protein interaction modulator，PPIM）**的设计与开发。

目前越来越多的PPIM已经上市或进入临床试验阶段。例如：BCL-2抑制剂venetoclax（ABT-199)被批准用于治疗慢性淋巴细胞白血病；LFA-1/ICAM-1抑制剂lifitegrast（SAR-1118）被批准用于治疗干眼症。

PPIM的研究和发现仍面临巨大挑战，PPI具有宽泛、普遍、无显著特征等特性，用于筛选传统靶点的化合物库不适合筛选PPIM。随着结构生物学和生物化学的发展，出现了专门的PPIM数据库，方便适用机器学习的方法进行筛选。

基于机器学习的传统方法

01

过度依赖人工提取和选择分子描述符和指纹特征，需要大量专业知识和人工干预

基于原子级GNN的方法

02

基于原子级图的GNN通常忽略了分子的层次信息，而药物中的官能团等分子层次信息才是决定药物与靶点相互作用的关键因素

作者研究

03

提出了一种基于分层图神经网络的分子表征学习框架，称之为HiGPPIM

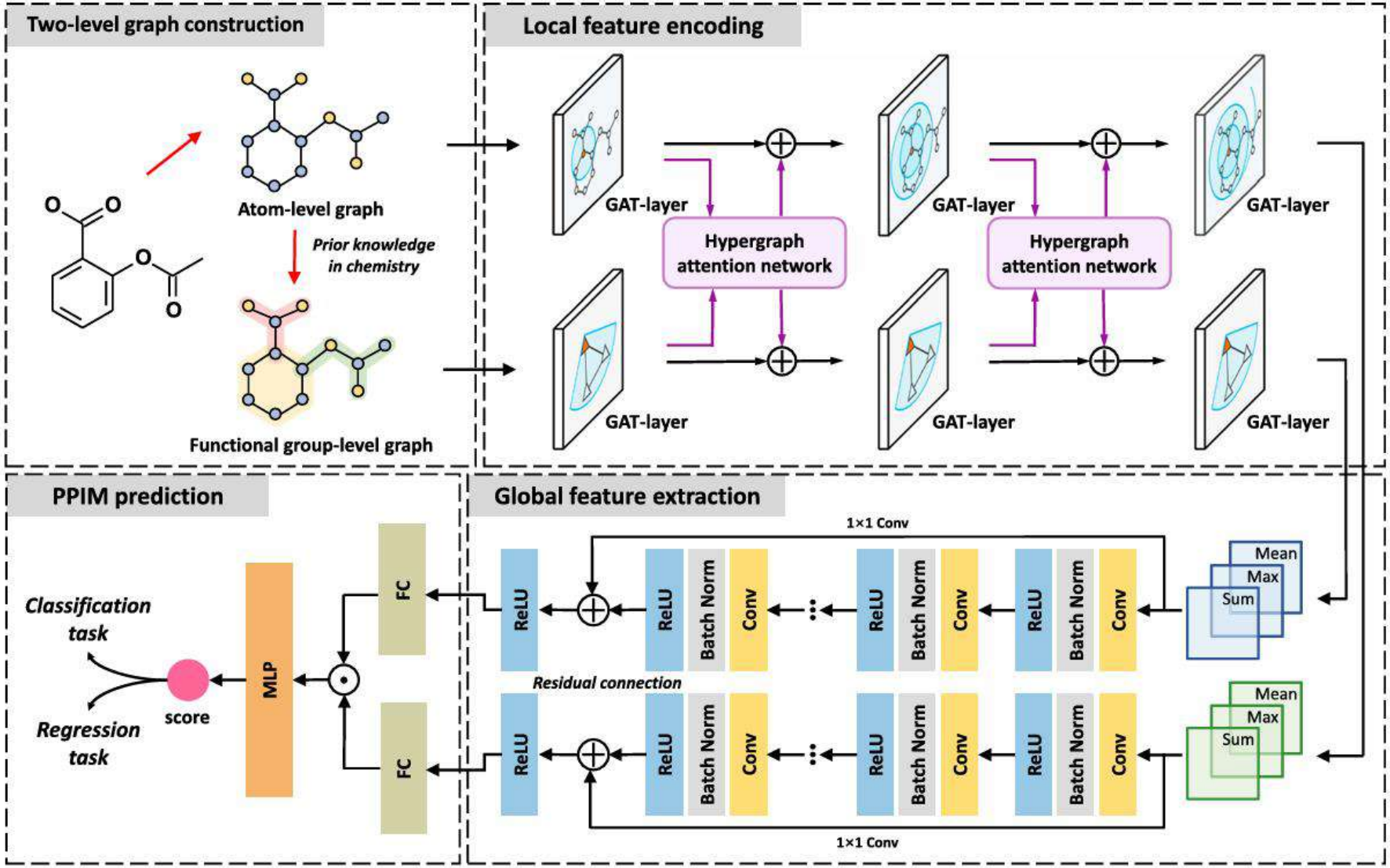
METHODS------（1）HiGPPIM的结构

一、两级图的构建

二、局部特征编码

三、全局特征提取

四、PPIM预测



METHODS------（2）两级图的构建

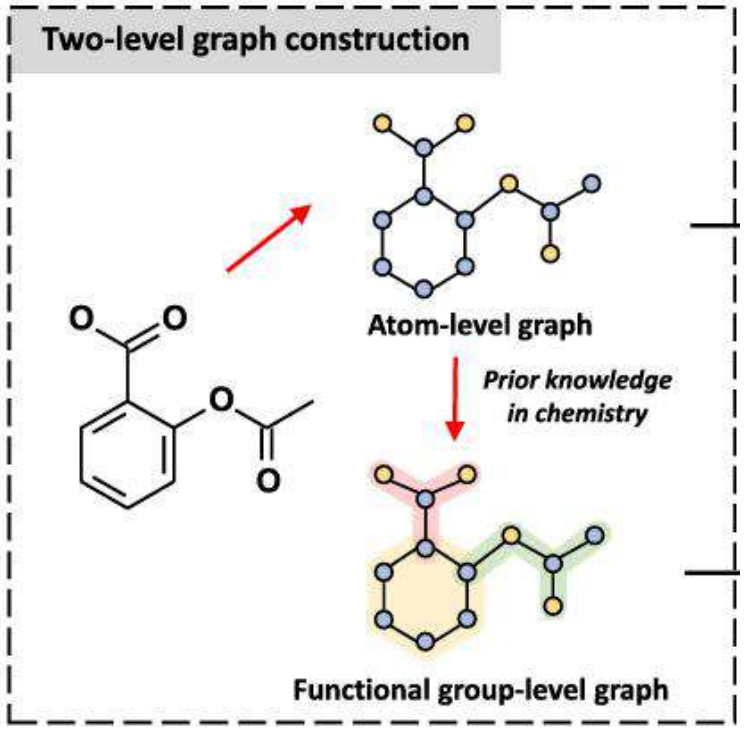
原子级图：分子可自然地用图来表示，以原子为节点、以化学键为边

$$G_{\text{atom}} = (V_{\text{atom}}, E_{\text{atom}})$$

官能团级图：在原子级图的基础上，利用已知化学知识，构建粗略的官能团级图

$$G_{fg} = (V_{fg}, E_{fg})$$

分子特征化：对于原子级图，用10种和3种键性来描述原子的物理化学性质、结构性质和局部环境。对于官能团级图，主要观察所包含原子和化学键的性质



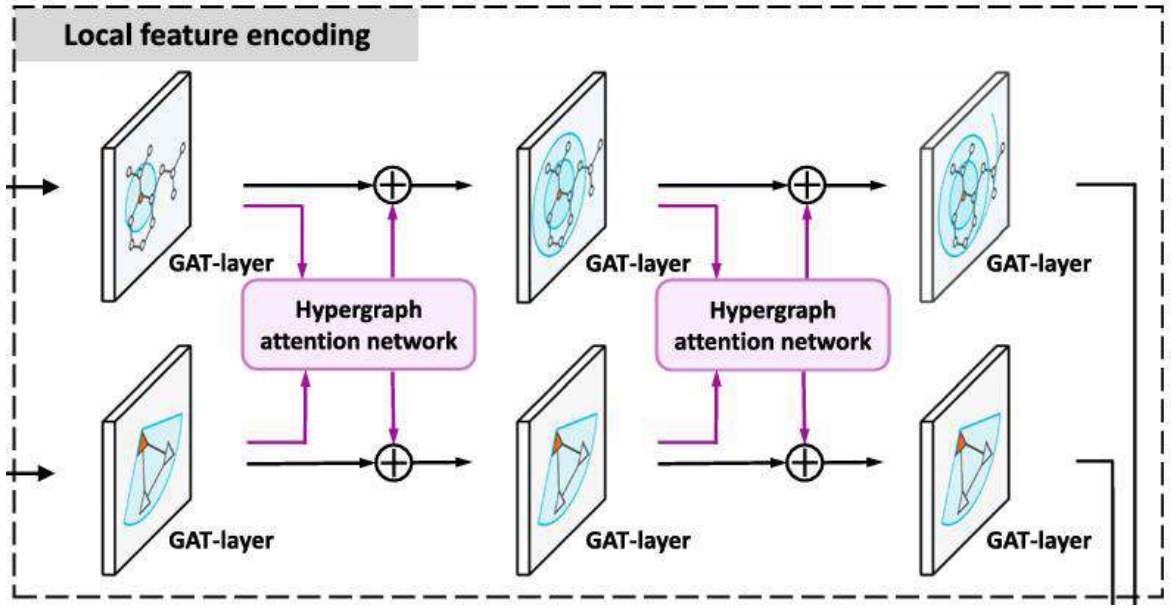
识别和定义官能团时，优先考虑环形结构和非环形结构：有环时选取最小环形为官能团；无环时先标记中心原子，将中心原子和与其通过化学键（单、双、三）直接相连的环境原子组合为官能团；最后定义上述规则未包含的官能团

METHODS------（3）局部特征编码

基于两级图结构，作者提出一种名为**“HiHyper”**的新型分层超图注意力网络。由两个级联模块组成：

一是**图注意力网络**，用于学习两张图中节点（即原子或官能团）的潜在表征；

二是**超图注意力网络**，作为中央枢纽，用于聚合和转换两种不同粒度的结构信息。



METHODS------（3）局部特征编码

图注意力网络：使用GAT分别从原子级图和官能团级图中提取分子特征。图注意力网络利用注意力机制，为节点邻居分配注意力权重，使得网络关注最相关的节点及其相互作用，从而提取出更有意义的特征。

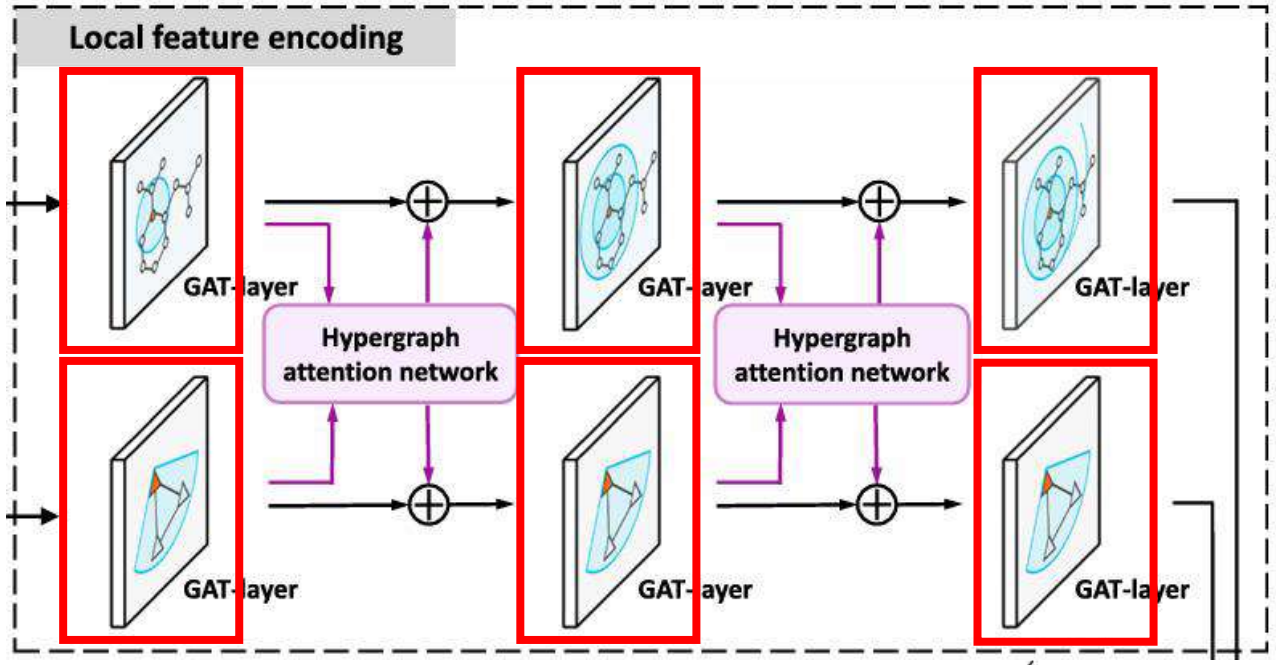
以原子级图为例，GAT的输入是一组原子特征 $\mathbf{X}_{\text{atom}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ $\mathbf{x}_i \in \mathbb{R}^M$

其中，*N*是原子个数，*M*是原子特征个数，第*j*个节点对第*i*个节点的注意力得分为：

$$\alpha_{i,j} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top \left[\mathbf{W}\mathbf{x}_i \| \mathbf{W}\mathbf{x}_j \| \mathbf{W}_e\mathbf{e}_{i,j}\right]\right)\right)}{\sum_{v_k \in \mathcal{N}(v_i)} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top \left[\mathbf{W}\mathbf{x}_i \| \mathbf{W}\mathbf{x}_k \| \mathbf{W}_e\mathbf{e}_{i,k}\right]\right)\right)}$$

其中，$\mathcal{N}(v_i)$ 是第*i*个节点的相邻节点，$\mathbf{e}_{i,j} \in \mathbb{R}^C$ 是*i*、*j*两个节点间边的特征，$\|$ 是连接操作，$\mathbf{W} \in \mathbb{R}^{M \times M'}$ $\mathbf{W}_e \in \mathbb{R}^{C \times M'}$ 是权重矩阵，$\mathbf{a} \in \mathbb{R}^{3 \times M'}$ 是单层前馈神经网络的权重向量，其中 *C* 是边缘特征。

METHODS------（3）局部特征编码



本实验中采取了多重注意力机制，新的节点特征描述为：$\mathbf{x}_i' = \Big\|_{k=1}^{K} \sigma\left(\alpha_{i,i}^k \mathbf{W}^k \mathbf{x}_i + \sum_{v_j \in \mathcal{N}(v_i)} \alpha_{i,j}^k \mathbf{W}^k \mathbf{x}_j\right)$

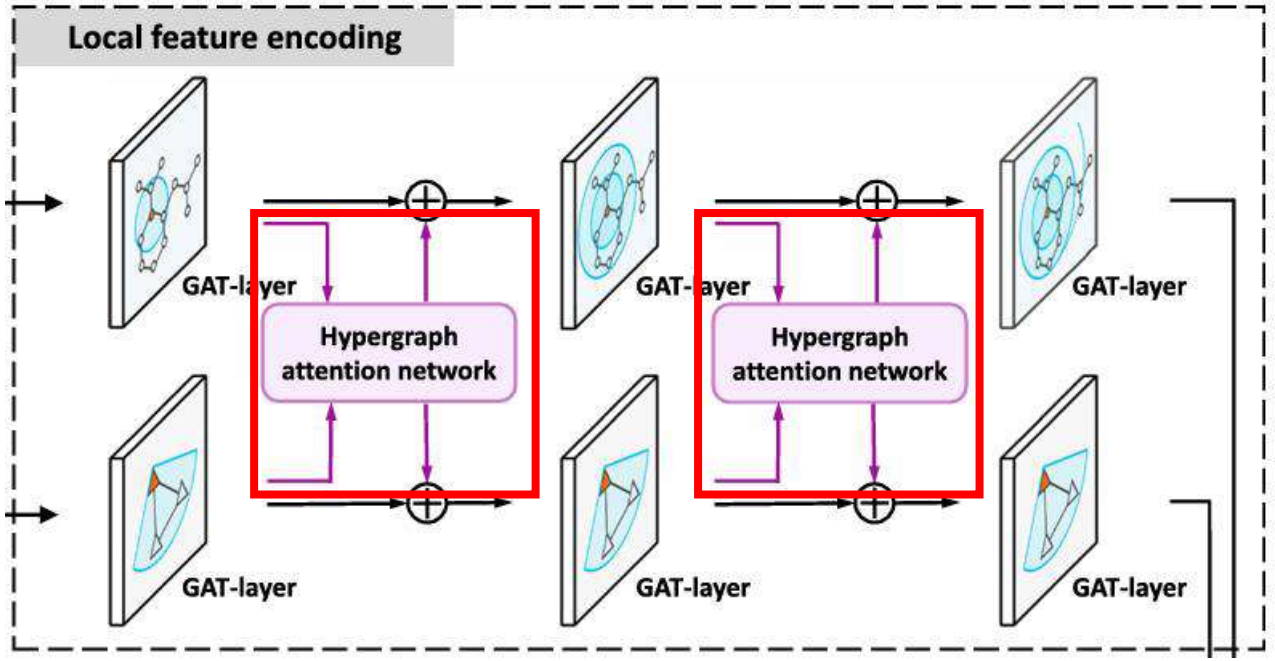
其中，σ 为*LeakyReLU*激活函数，$\alpha_{i,j}^k$ 和 $\mathbf{W}^k$ 分别是第k个注意力机制的注意力得分和注意力权重矩阵

METHODS------（3）局部特征编码

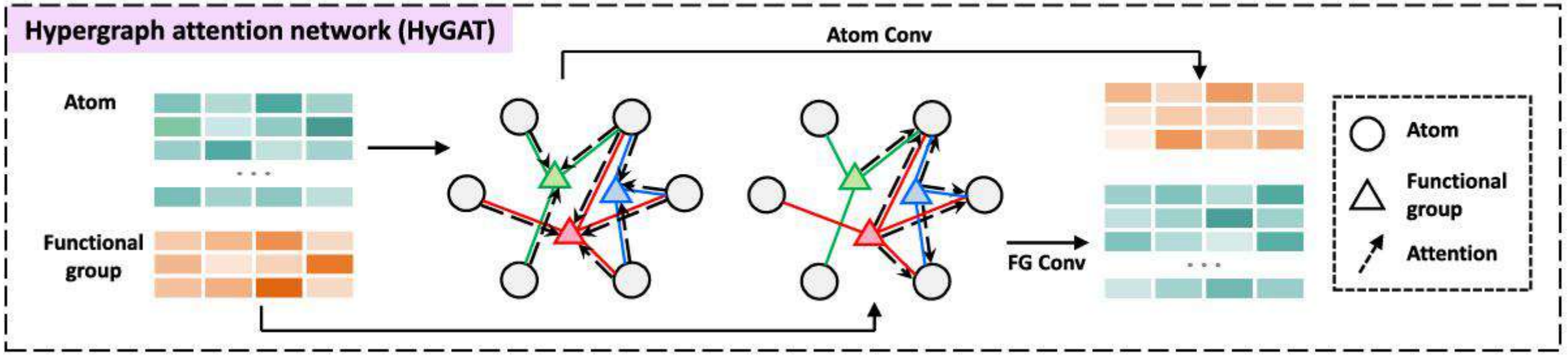
超图注意力网络：通过使用GAT，可得出在原子级和官能团级不同粒度下的分子结构特征，但并未捕捉原子和官能团之间的相关性。为此进一步构建了原子-官能团（ATOM-FG)超图 $G_h = (V_h, E_h)$

本实验设计了超图注意力网络（**HyGAT**）。

在分子中，每个官能团会包含多个原子，每个原子也会包含在多个官能团中。然而，不同的原子（官能团）对与其相关的官能团（原子）的重要性因其位置和电负性等因素而异。因此，聚合节点（原子）和超边（官能团）特征的过程中，分别对入射矩阵***H***引入注意力机制，以区分原子和官能团的贡献。



METHODS------（3）局部特征编码



超图卷积分为原子卷积和官能团卷积两部分，采用双重注意力机制学习原子级和官能团级分子特征。

原子卷积将GAT产生的原子级特征和相关官能团聚合在一起，产生具有原子特征的官能团级特征。同样地，官能团卷积将GAT产生的官能团级特征和相关原子聚合在一起，产生具有官能团特征的原子级特征。

$$\mathbf{X}''_{fg} = \sigma\left(\mathbf{B}^{-1}\mathbf{H}^{\top}\mathbf{X}'_{\text{atom}}\mathbf{\Theta}_{\text{atom}}\right) \qquad \mathbf{X}''_{\text{atom}} = \sigma\left(\mathbf{D}^{-1}\mathbf{H}\mathbf{X}'_{fg}\mathbf{\Theta}_{fg}\right)$$

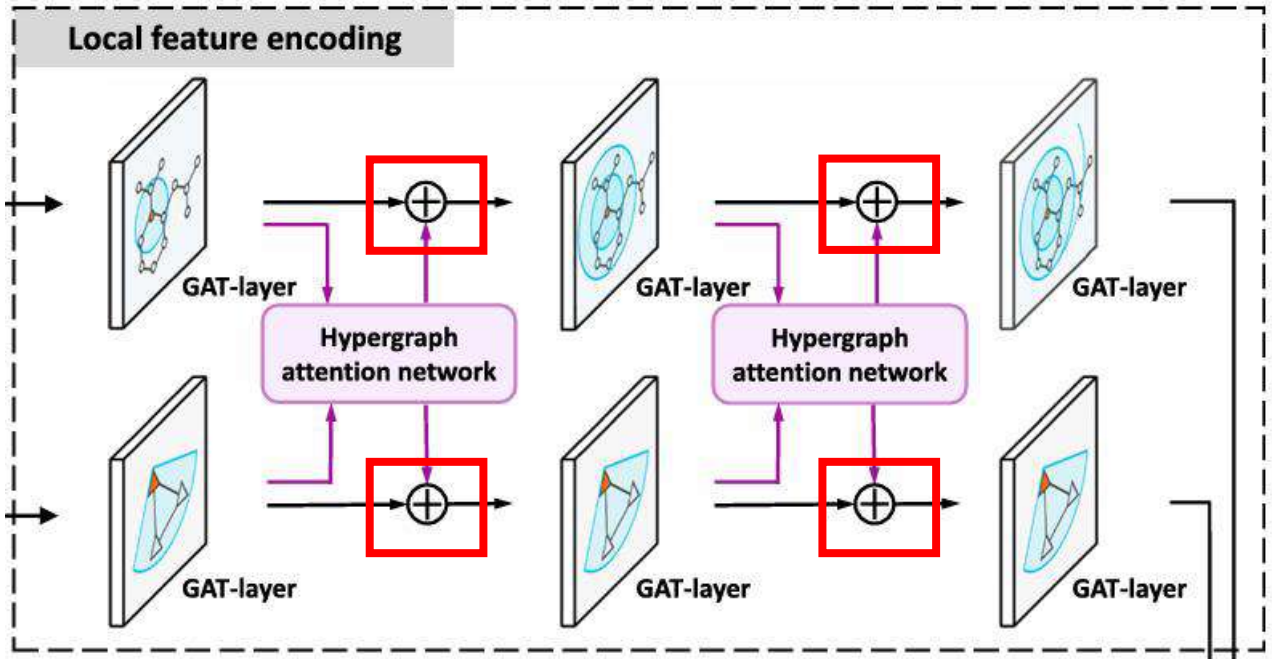
METHODS------（3）局部特征编码

HiHyper的级联学习：通过结合GAT和HyGAT，将HiHyper层定义为如下：

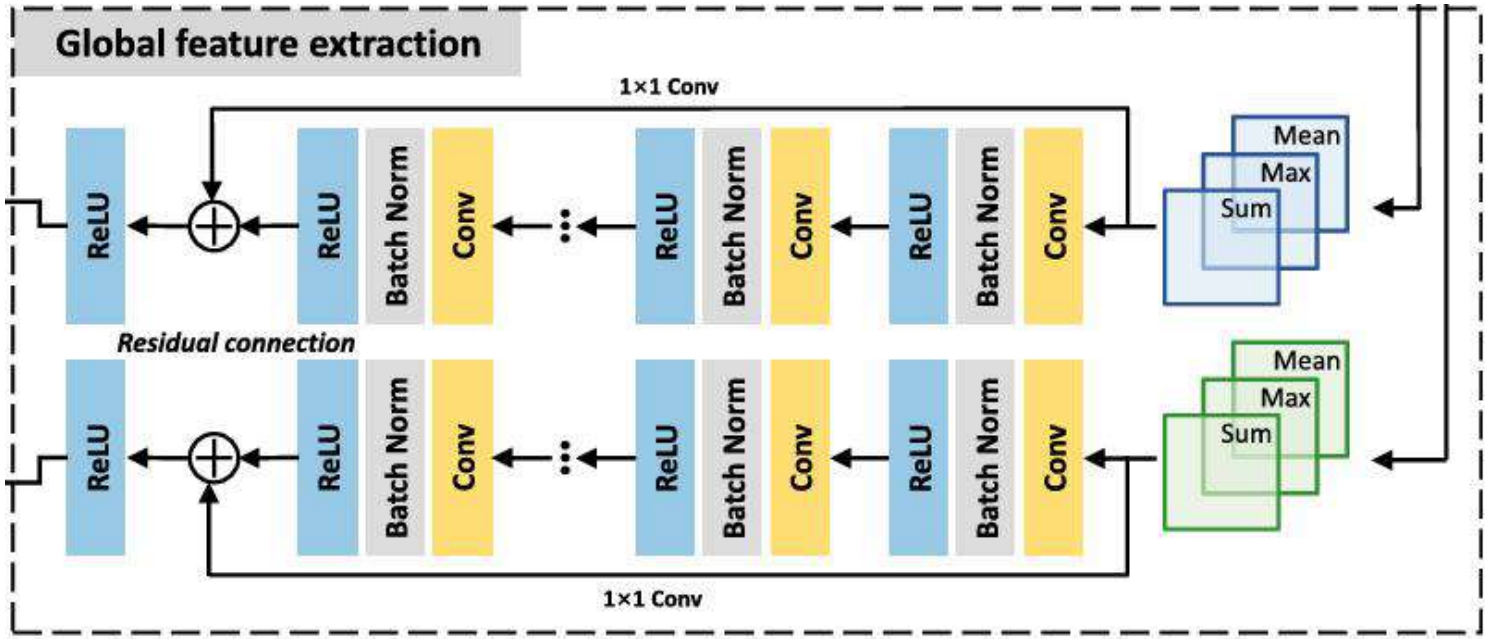
$$\mathbf{X}_{\text{atom}}^{l} = GAT_{\text{atom}}\left(\mathbf{X}_{\text{atom}}^{l-1}\right) + HyGAT_{FgConv}\left(GAT_{fg}\left(\mathbf{X}_{fg}^{l-1}\right)\right)$$

$$\mathbf{X}_{fg}^{l} = GAT_{fg}\left(\mathbf{X}_{fg}^{l-1}\right) + HyGAT_{\text{AtomConv}}\left(GAT_{\text{atom}}\left(\mathbf{X}_{\text{atom}}^{l-1}\right)\right)$$

将GAT获得的原子级特征和HyGAT聚合官能团后得到的原子级特征进行残差连接，得到更为全面的原子级特征。官能团级特征亦是如此。



METHODS------（4）全局特征提取



在HiHyper模块后，用池化层和卷积块来提取分子的全局特征。堆叠多个卷积块来提取深层次特征信息，并通过1×1的卷积层引入残差连接加快训练过程。

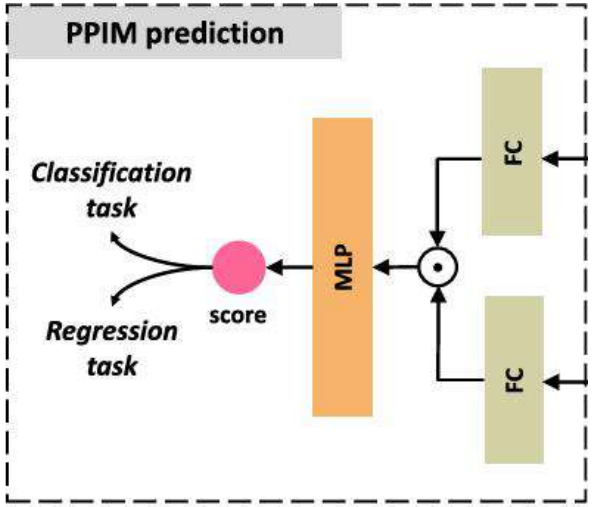
$$\mathbf{X}_{\text{atom}}^{pool} = \mathbf{X}_{\text{atom}}^{\text{pool}_{\text{sum}}} \,||\, \mathbf{X}_{\text{atom}}^{pool_{mean}} \,||\, \mathbf{X}_{\text{atom}}^{\text{pool}_{\text{max}}},$$

$$\mathbf{X}_{\text{atom}}^{conv} = \sigma\left(BN\left(Conv\left(\mathbf{X}_{\text{atom}}^{pool}\right)\right)\right).$$

$$\mathbf{X}_{\text{atom}}^{\text{out}} = \sigma\left(\mathbf{X}_{\text{atom}}^{conv_n} + Conv_{1\times1}\left(\mathbf{X}_{\text{atom}}^{conv_n}\right)\right)$$

METHODS------（5）PPIM预测

在得到原子级全局特征和官能团级全局特征后，应用全连接层FC投射到新的嵌入空间并计算Hadamard乘积，原子级和官能团级分子表征被输入到MLP中完成预测任务。



在PPIM识别任务中，运用二元交叉熵作为损失函数：

$$\mathcal{L}oss_{cls} = -\frac{1}{\Gamma}\sum_{i=1}^{\Gamma}[y_i \log \sigma(\hat{y}_i) + (1-y_i)\log(1-\sigma(\hat{y}_i))]$$

在PPIM药物预测任务中，运用均方误差作为损失函数：

$$\mathcal{L}oss_{reg} = \frac{1}{\Gamma}\sum_{i=1}^{\Gamma}(y_i - \hat{y}_i)^2$$

评价指标

对于分类任务，采取准确率（ACC）、Matthews 相关系数（MCC）、F1分数、ROC曲线下面积（AUC）作为评价指标。

对于回归任务，采取平均绝对误差（MAE）、均方根误差（RMSE）、皮尔逊相关系数（ρ）、肯德尔相关系数（τ）、斯皮尔曼相关系数（rs）作为评价指标。

对照基准方法

SMMPPI：采用拓展连通性指纹（ECFPs）和随机森林算法构建PPI分类器

PPI-ML：基于随机树、逻辑回归、ECFPs训练过的支持向量机的集成学习模型

SELPPI：堆叠集成计算框架，包含6个基于树的机器学习模型，使用7种特征描述符和遗传算法训练

PPIM识别

PPI Family	Method	MCC	F1 score	AUC	ACC
Bcl2-Like/Bak-Bax	HiGPPIM	**0.953±0.000**	0.976±0.000	**0.997±0.000**	**0.976±0.000**
	SELPPI	0.945±0.018	0.970±0.010	0.984±0.006	0.971±0.010
	PPI-ML	**0.953**	**0.977**	0.976	**0.976**
	SMMPPI	0.858	0.927	0.993	0.929
Bromodomain/Histone	HiGPPIM	**0.878±0.009**	**0.939±0.004**	**0.981±0.003**	**0.939±0.004**
	SELPPI	0.791±0.020	0.898±0.010	0.960±0.006	0.894±0.010
	PPI-ML	0.780	0.891	0.890	0.890
	SMMPPI	0.780	0.892	0.966	0.890

文章提出的HiGPPIM在所有数据集上均表现出较为优秀的识别能力，各项数据均处于最高水平，在一部分数据集上甚至可以实现准确率ACC提升4.5%-8.9%。

HiGPPIM在AUC指标方面具有显著优势，各数据集上平均值为0.992，体现了HiGPPIM识别PPIM的强大能力。

PPIM药物预测

PPI Family	Method	ρ	τ	r_s	RMSE	MAE
Bcl2-Like/Bak-Bax	HiGPPIM	**0.547±0.020**	0.486±0.037	**0.663±0.041**	0.954±0.022	**0.759±0.043**
	SELPPI	0.427±0.073	**0.580±0.082**	0.579±0.068	**0.933±0.045**	0.800±0.044
Bromodomain/Histone	HiGPPIM	**0.839±0.024**	**0.736±0.029**	**0.849±0.024**	**0.610±0.021**	**0.449±0.012**
	SELPPI	0.416±0.070	0.618±0.067	0.636±0.080	0.969±0.072	0.793±0.064
CD4/gp120	HiGPPIM	**0.863±0.021**	**0.618±0.047**	**0.715±0.060**	**0.768±0.077**	**0.642±0.054**
	SELPPI	-	-	-	-	-
LEDGF/IN	HiGPPIM	**0.558±0.013**	0.420±0.059	0.481±0.068	0.837±0.014	0.694±0.009
	SELPPI	0.478±0.111	**0.631±0.124**	**0.790±0.049**	**0.654±0.027**	**0.609±0.028**

三种基准模型，只有SELPPI提供了有效的PPIM预测模型，与其相比，HiGPPIM几乎在所有数据集上表现出良好的性能，但在LEDGF/IN数据集上表现较差。且在所有数据集上， HiGPPIM的皮尔逊相关系数均为最高，表面着预测情况与实况有很强的线性相关性。

总的来说HiGPPIM在PPIM识别和药物预测方面能表现出基本优于、少数持平现有手段的性能。

BIT

计算机科学与技术前沿——生信方向专题汇报

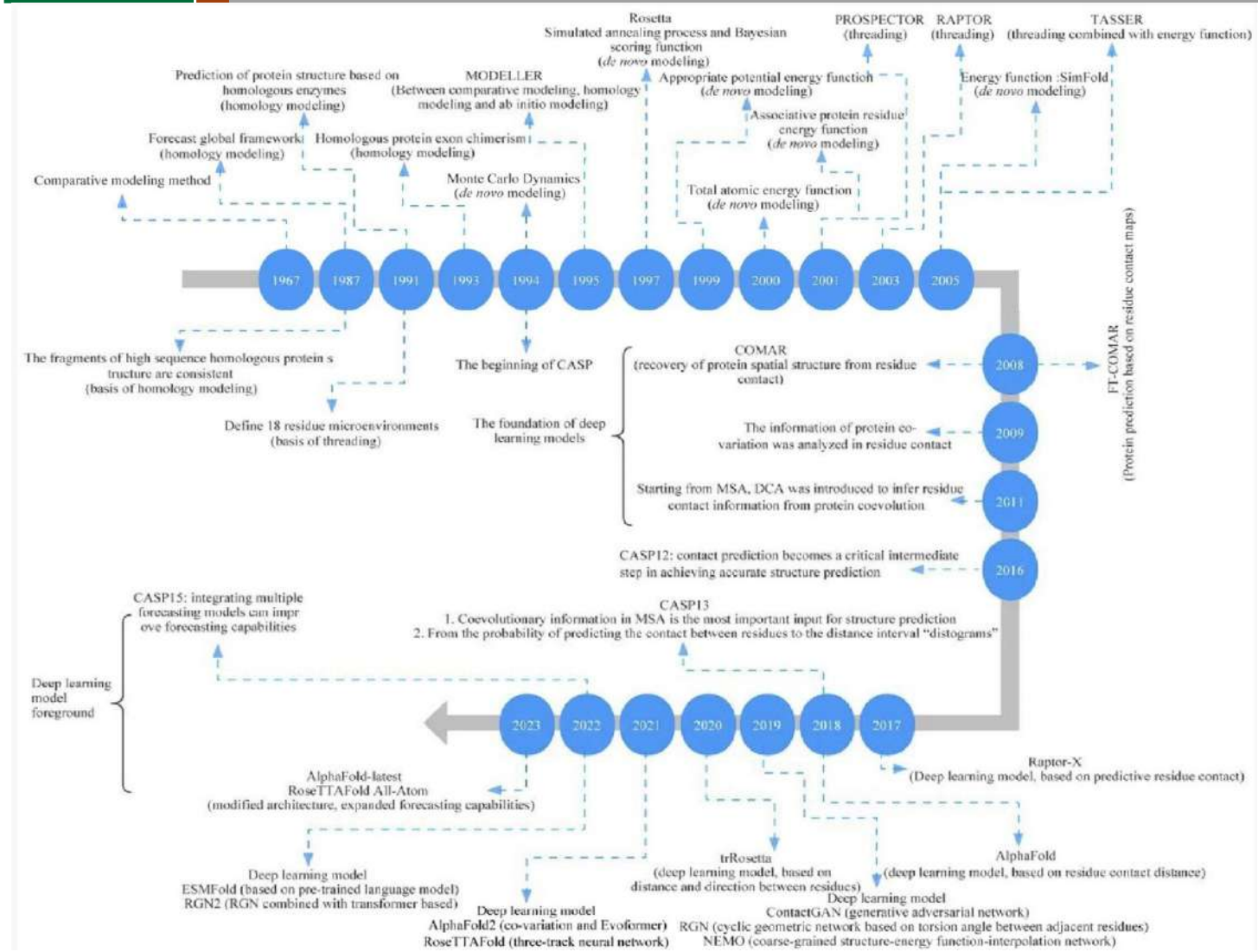
蛋白质结构预测模型 AlphaFold2

汇报人：张洪洋　　时间：2024-11-6

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

德以明理　学以精工

蛋白质结构预测方法发展历程



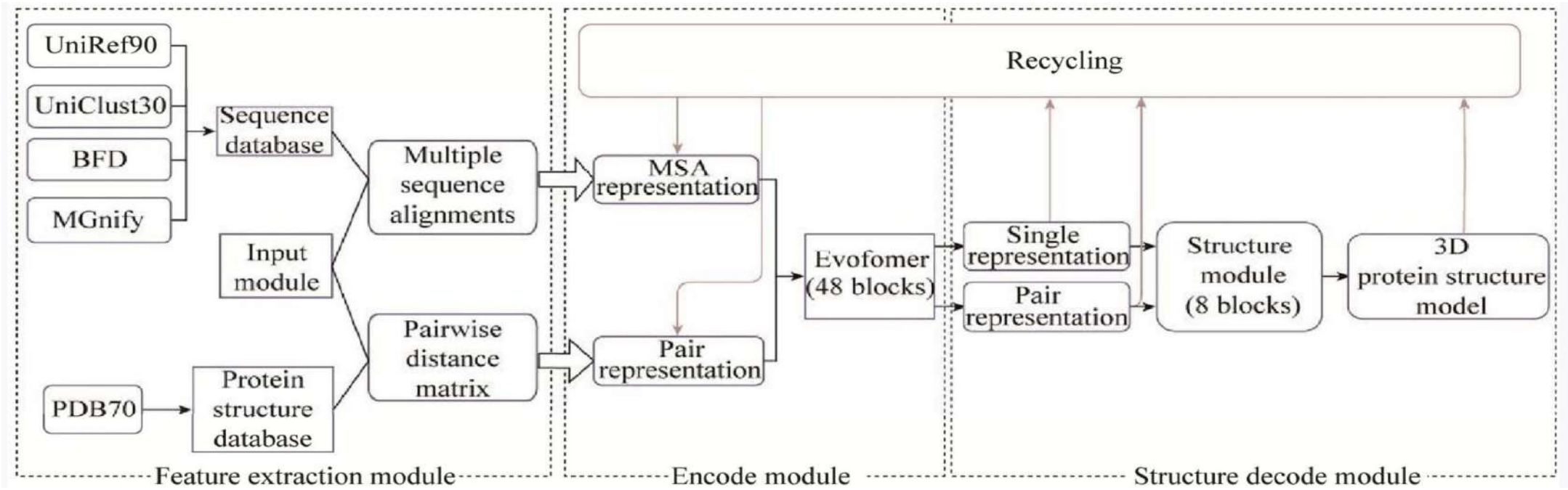
Alpha Fold2

Rose TTAFold

ESMFold

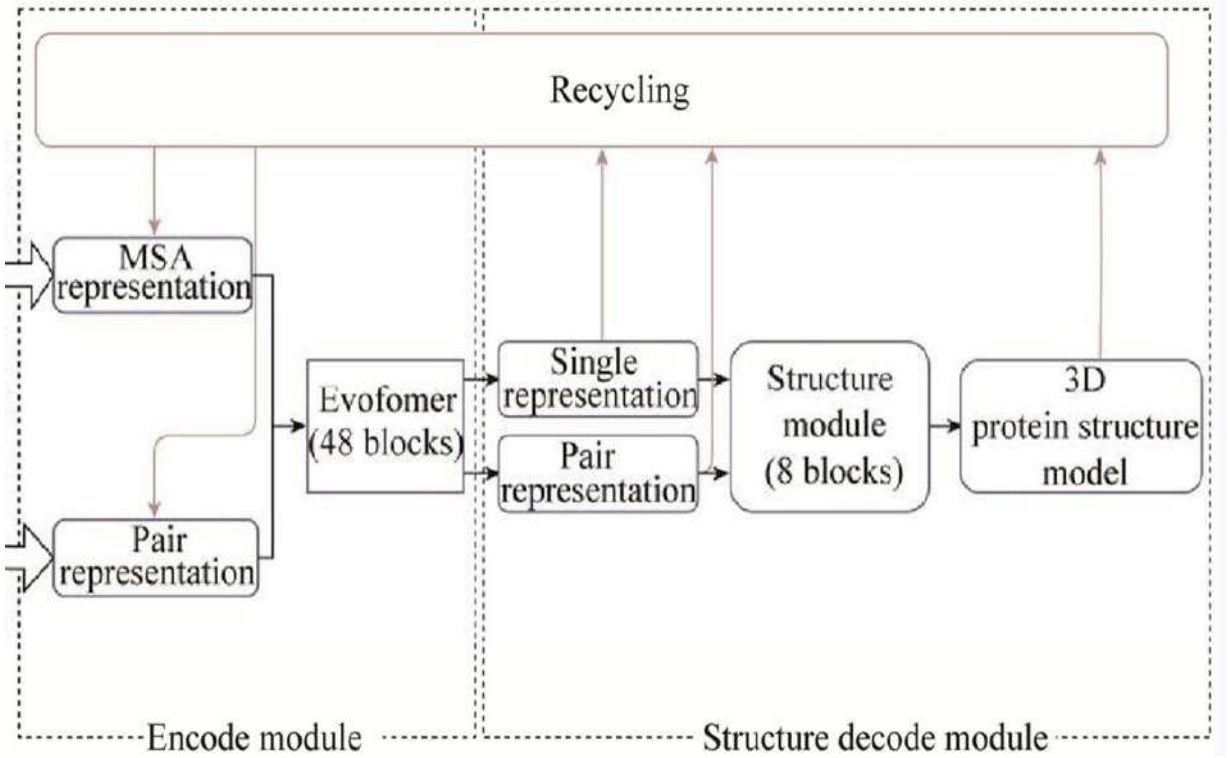
总体框架

AF2总体框架如图所示，分为特征提取模块、编码模块、结构解码模块。输入模块根据给定的氨基酸序列，在序列数据库中寻找其同源序列，并进行多序列比对。MSA可以反映出蛋白质序列之间的相似性和共进化信息，这些信息对于预测蛋白质结构尤为重要。同时输入模块检查是否有任何同源序列存在已知的三维结构，并在蛋白质结构数据库中查找；如有，输入模块会构建一个两两距离矩阵，表示每一对氨基酸之间的空间距离。随后输入模块生成MSA表示和成对表示，其中MSA表示是一个三维矩阵，表示每个氨基酸在MSA中的位置、频率和共进化信息；成对表示也是三维矩阵，表示氨基酸之间结构约束信息特征，包括每一对氨基酸之间的距离、角度和相互作用



Alph Fold2总体框架

核心子模块



Recycling

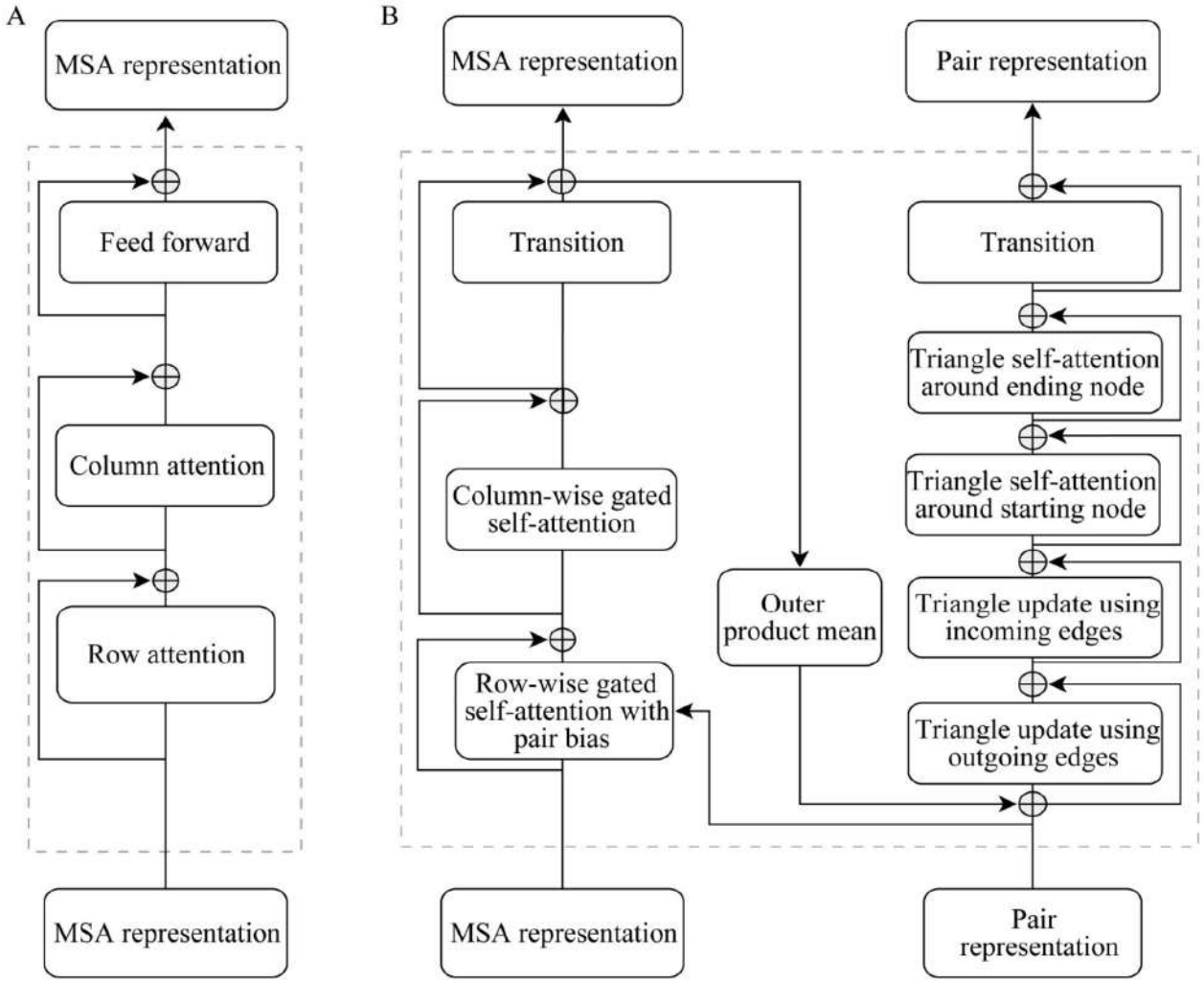
编码模块

Evoformer模块能够利用MSA和残基对之间的共进化信息来推理蛋白质的空间和进化关系。AF2使用48个不共享权重的Evoformer模块，每个模块有一个MSA表示和一个成对表示作为输入，并输出更新后的MSA表示和成对表示。每个Evoformer模块通过三角自注意力机制和几何变换来更新这两个表示

结构解码模块

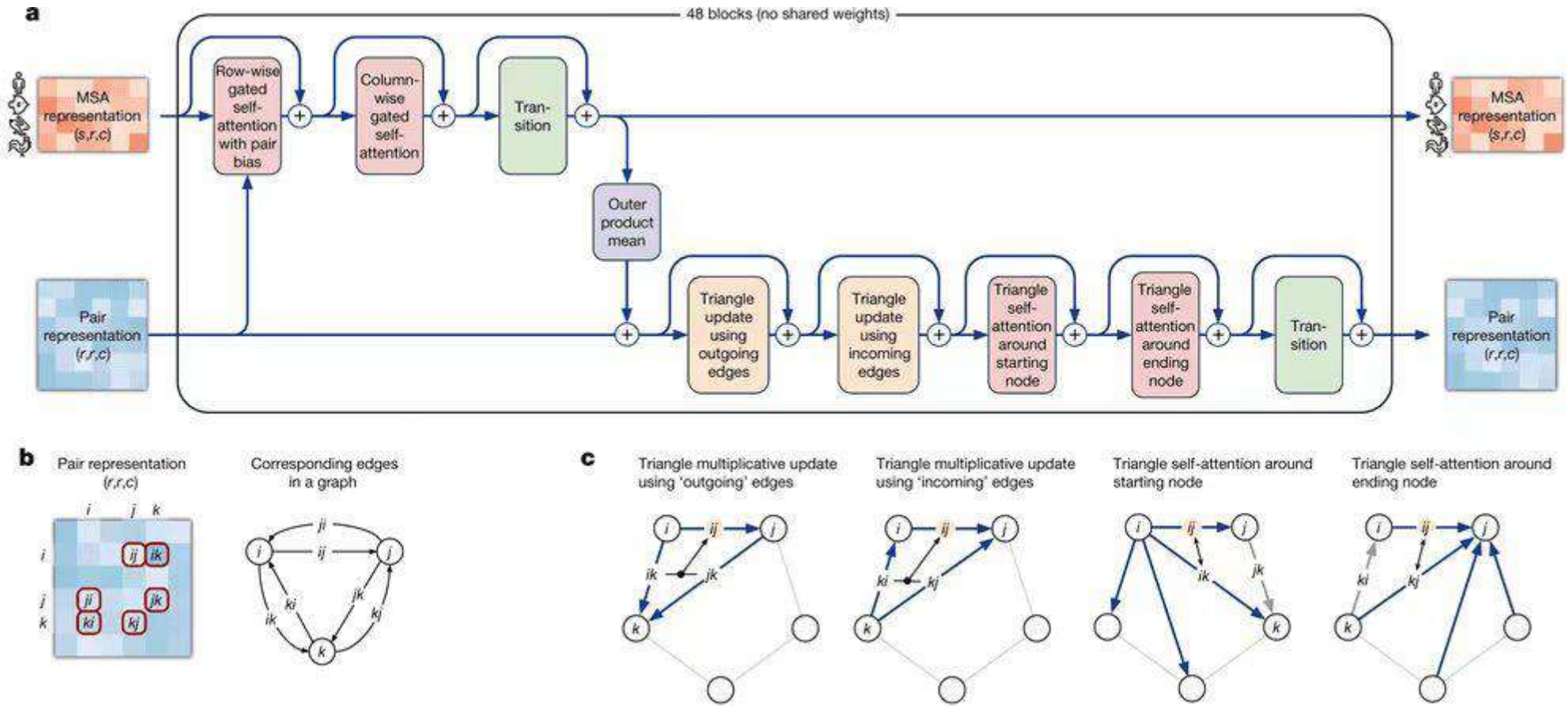
可以将编码模块的输出结果转换为目标蛋白质的三维结构。结构解码模块的关键组件是不变点注意力（IPA），它是一种几何感知的注意力机制，用于更新单一表示。IPA操作的最终注意力值在三维空间中是等变的，即不管蛋白质结构在三维空间中如何变换，不变点注意力都能保持相同的输出，这有助于提高蛋白质结构预测的准确性和稳定性

Evoformer的开发灵感来源于MSA-Transformer。Transformer是一种新兴的自注意力模型，利用自注意力机制来提取序列数据的内在特征，具有广泛的人工智能应用潜力。在Transformer的基础上延伸出的MSA-Transformer使用MSA表示(MSA representation)作为输入，通过注意力(attention)机制来处理蛋白质序列的信息。Evoformer使用了2组类似MSA-Transformer的结构，分别用于捕捉氨基酸残基间的多序列比对信息和结构约束信息特征，从而提高了预测质量

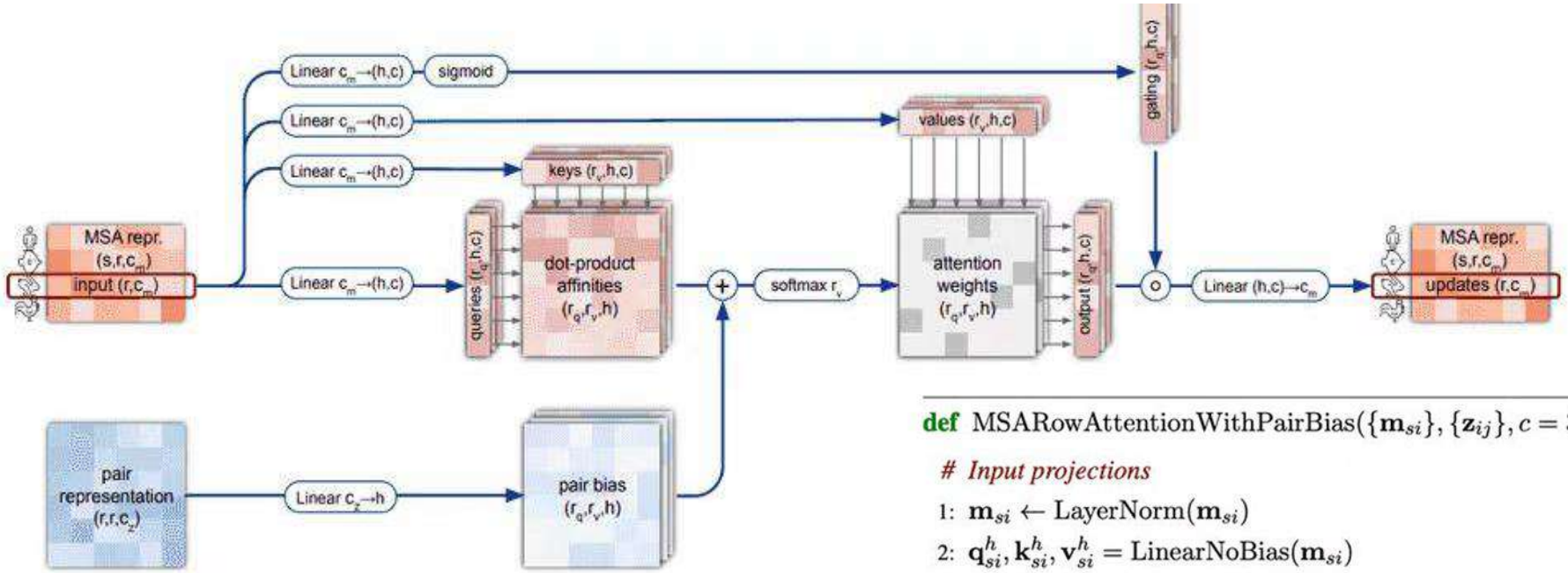


MSA-Transformer与Evoformer架构对比图

编码模块


a
48 blocks (no shared weights)
MSA representation (s,r,c)
Row-wise gated self-attention with pair bias
Column-wise gated self-attention
Tran-sition
Outer product mean
Pair representation (r,r,c)
Triangle update using outgoing edges
Triangle update using incoming edges
Triangle self-attention around starting node
Triangle self-attention around ending node
Tran-sition
MSA representation (s,r,c)
Pair representation (r,r,c)
b
Pair representation (r,r,c)
Corresponding edges in a graph
c
Triangle multiplicative update using 'outgoing' edges
Triangle multiplicative update using 'incoming' edges
Triangle self-attention around starting node
Triangle self-attention around ending node

ROW-WISE GATED SELF-ATTENTION



def MSARowAttentionWithPairBias($\{\mathbf{m}_{si}\}, \{\mathbf{z}_{ij}\}, c = 32, N_{\text{head}} = 8$) :

 # Input projections

$\mathbf{m}_{si} \leftarrow \text{LayerNorm}(\mathbf{m}_{si})$

$\mathbf{q}_{si}^h, \mathbf{k}_{si}^h, \mathbf{v}_{si}^h = \text{LinearNoBias}(\mathbf{m}_{si})$ $\quad \mathbf{q}_{si}^h, \mathbf{k}_{si}^h, \mathbf{v}_{si}^h \in \mathbb{R}^c,\ h \in \{1, \ldots, N_{\text{head}}\}$

$b_{ij}^h = \text{LinearNoBias}(\text{LayerNorm}(\mathbf{z}_{ij}))$

$\mathbf{g}_{si}^h = \text{sigmoid}\left(\text{Linear}(\mathbf{m}_{si})\right)$ $\quad \mathbf{g}_{si}^h \in \mathbb{R}^c$

 # Attention

$a_{sij}^h = \text{softmax}_j\left(\frac{1}{\sqrt{c}}\, \mathbf{q}_{si}^{h\top} \mathbf{k}_{sj}^h + b_{ij}^h\right)$

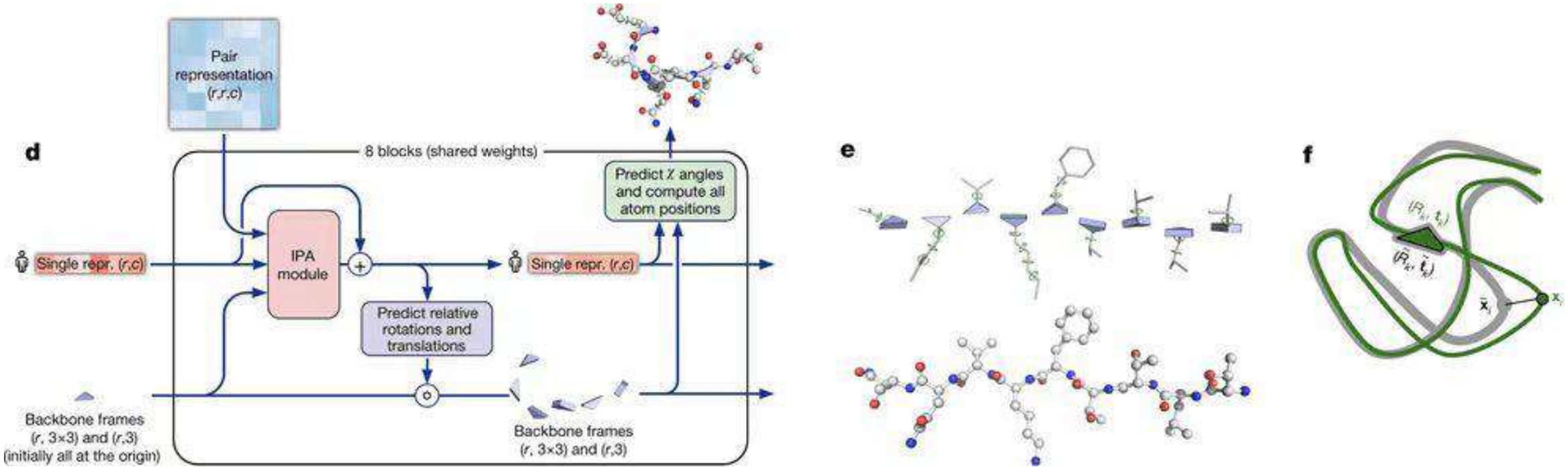
$\mathbf{o}_{si}^h = \mathbf{g}_{si}^h \odot \sum_j a_{sij}^h \mathbf{v}_{sj}^h$

 # Output projection

$\tilde{\mathbf{m}}_{si} = \text{Linear}\left(\text{concat}_h(\mathbf{o}_{si}^h)\right)$ $\quad \tilde{\mathbf{m}}_{si} \in \mathbb{R}^{c_m}$

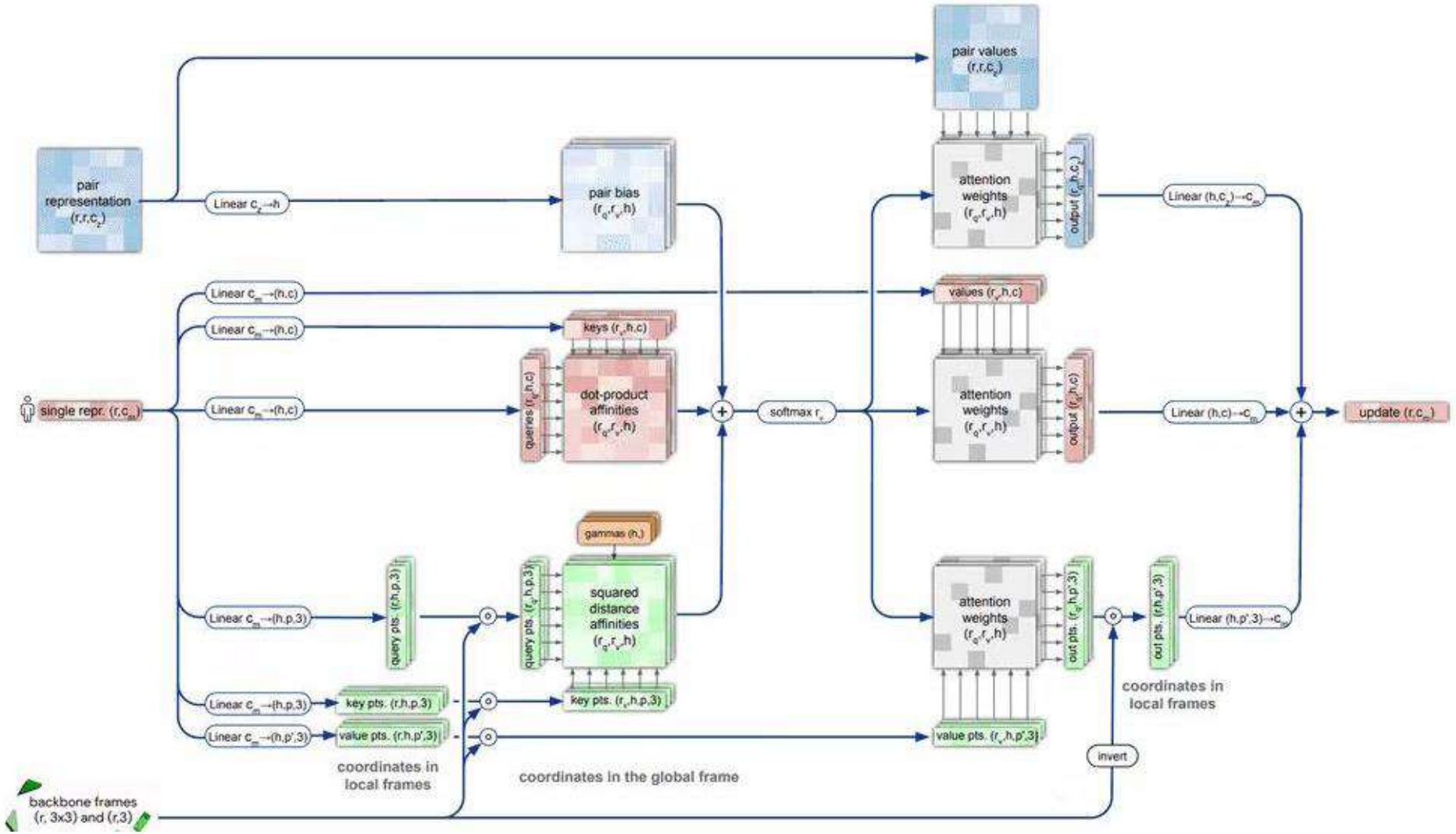
return $\{\tilde{\mathbf{m}}_{si}\}$

结构解码模块



Structural module 使用原始的 sequence embedding 和 pair representation 来预测结构。

不变点注意板块(IPA)


pair representation
Linear
pair bias
pair values
attention weights
output
Linear
single repr.
Linear
Linear
keys
dot-product affinities
queries
softmax
values
attention weights
output
Linear
update
gammas
Linear
query pts.
squared distance affinities
key pts.
attention weights
out pts.
Linear
Linear
key pts.
Linear
value pts.
value pts.
coordinates in local frames
coordinates in the global frame
invert
coordinates in local frames
backbone frames

总结

- Alpha Fold2作为基于深度学习的蛋白质结构预测模型，通过独特的原理和架构，实现了高准确度的快速蛋白质结果预测，并在生物学和医学的研究中发挥多方面的作用。但是AF2还存在严苛的算力需求的缺陷，难以开展大规模应用，并且在某些蛋白的结构预测应用上存在着结构误差，有待相关技术工作者对其进行改良以满足更广泛的应用场景。目前基于深度学习开发的模型在生物学领域有着广阔的应用前景，不仅用于预测蛋白质结构，还为生物学各方面研究提供了工具。

感谢观看

小组成员：陈籽旭 刘天依 余宏骏 乔江洋 张洪洋